

Comprehensive experimental design for chemical engineering processes: a two - layer iterative design approach

Hui Yu^a, Hong Yue^a, Peter Halling^b

^a*Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK*

^b*WestCHEM Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow G1 1XL, UK*

Abstract

A systematic framework for optimal experimental design (OED) of multiple experimental factors is proposed to support data collection in chemical engineering systems with the purpose to obtain the most informative data for modeling. The structural identifiability is firstly investigated through a combined procedure with the generating series method and the identifiability tableau. Next the parameter estimability is analyzed via the orthogonalized sensitivity analysis in order to identify crucial and identifiable model parameters. Traditionally OED treats separate problems such as the choice of input conditions, the selection of variables to measure, and the design of sampling time profile. A new OED strategy is proposed that optimizes these interdependent factors in one framework. An iterative two-layer design structure is developed. In the lower layer for observation design, the sampling profile and the measurement set selection are combined and formulated as a single integrated observation design problem, which is relaxed to a convex optimization problem that can be solved with a local method. Thus the measurement set

selection and the sampling profile can be determined simultaneously. In the upper layer for input design, the optimization of input intensities is obtained through stochastic global searching. In this way, the multi-factor optimization problem is solved through the integration of a stochastic method, for the upper layer, and a deterministic method, for the lower layer. Case studies are conducted on two biochemical systems with different complexities, one is an enzyme kinetically controlled synthesis system and the other one is a lab-scale enzymatic biodiesel production system. Numerical results demonstrate the effectiveness of this double-layer OED optimization strategy in reducing parameter estimation uncertainties compared with conventional approaches.

Keywords: optimal experimental design (OED), multi-factor optimization, input conditions, sampling time profile, measurement set selection, chemical reaction systems.

1. Introduction

Mathematical models are widely used in chemical and biochemical process engineering since the mathematical representation enables to reproduce real dynamic processes in a simulation environment (Baltes et al., 1994; van Riel, 2006; Bogacka et al., 2011; Villaverde et al., 2014). These models can be used to explore the underlying nature of specific reactions, to better understand the dynamics of individual components and their interactions, to control and predict the future behavior of systems and to test hypotheses (Phair, 1997; Peleg et al., 2002; Fages et al., 2004; de Brauwere et al., 2009; Liepe et al., 2013; Yu et al., 2015). A typical modeling procedure consists of several important steps (Franceschini and Macchietto, 2008), as shown in

Fig.1. Once one or several candidate models are proposed from prior knowledge, it is necessary to investigate if it's possible to obtain unique solutions for model parameters under ideal conditions of noise free observations and error-free model structures, if not, alternative models need to be proposed. For those structurally identifiable models, parameter sensitivity analysis and estimability analysis are required which will help to make model calibration more specific on those key parameters, whereas non-important parameters can be kept on their nominal values or even be removed so as to reduce the model complexity. The most suitable model can then be determined through fitting with experimental data, which is referred to as model calibration in Fig.1. The established model needs to be further validated using experimental data.

Model development of process systems is normally an iterative process that includes steps on data collection, model selection, model calibration and model validation until a satisfactory model is obtained with acceptable predictive capabilities. It requires large amounts of experimental data at all modeling steps. For chemical reaction systems, a typical method is to represent reactions into a set of coupled differential equations based on certain conceptual framework, e.g. mass-action laws. The reactants and products involved in the reaction network are therefore interconnected with kinetic parameters, whose values are generally unknown *a priori*. One of the main goals in model building is then to estimate those unknown parameters based on experimental data. However, measurement of process variables especially reactants is restricted by many factors such as sensor technology, operation constraints, limited time and budget, etc. Constraints on inputs can also

74 affect implementation of experiments. What's more measurement data are
 75 inevitably contaminated with experimental noise. The lack of sufficient and
 76 accurate measurement data makes model development a challenging task
 77 especially when the system is high dimensional, nonlinear with poorly un-
 78 derstood dynamics like many complex biological or biochemical networks.

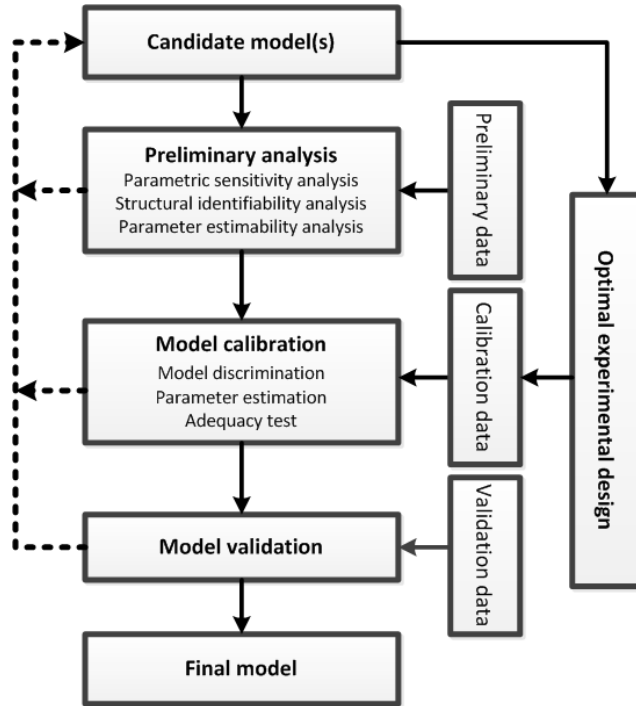


Figure 1: Data-based model building process

79 In data-based model development, it is essential to obtain high quality
 80 and informative measurement data with less experimental efforts if possi-
 81 ble. Therefore, modern experimental design techniques play important roles
 82 in model building process at various stages. The purpose of optimal ex-
 83 perimental design (OED) is to devise necessary experiments that are most

likely to generate data that will best facilitate the identification of model structure and the determination of model parameters (Faller et al., 2003). Typically an OED problem can be formulated as a dynamic optimization problem with respect to the design factors of interest. The major objective of OED is to maximize the data information through a measure of certain scalar function of Fisher information matrix (FIM) (Balsa-Canto et al., 2008). The design factors can normally be classified into two groups, i.e., the input design factors and the observation design factors. The former determines the stimulation and control actions, e.g., the initial conditions, the external time-dependent input conditions. These input factors will change the system dynamics. The latter is to determine which to measure, when to measure and where to measure, for example, design of sampling time profiles and design of measurement set selection. Here the measurement set refers to the choice of variables to be measured.

Various OED methods have been developed for data-based modeling of chemical, biological and wider systems aiming at individual experimental factors such as a factor in input settings (Chianeh et al., 2011; Yue et al., 2013) or a factor in observation design (Kutalik et al., 2004; Brown et al., 2008; Asyali, 2010; He et al., 2010). When a single factor is determined individually through OED, the design result and the overall information contained in the experimental data are dependent on other factors that are not included in the design. If those non-designed factors are not properly chosen, the quality of the experimental data cannot be guaranteed. For a dynamic system to be modeled, such dependence among experimental factors exist between the input factors, the measurement factors, and the interaction between them. To

109 reduce the uncertainty in single factor design and increase the data quality,
110 a more effective OED should support optimization of multiple experimental
111 factors in a systematic way. Very few works have been reported on how to
112 tackle OED of multiple experimental design factors mainly because it is very
113 difficult to obtain OED solutions for multiple design variables for a complex
114 nonlinear system, not to mention the system constraints and the operational
115 constraints that need to be considered in optimization. One option for multi-
116 factor OED is to implement the optimization of multiple experimental pa-
117 rameters through a sequential design strategy in which each single factor is
118 designed iteratively and the interested experimental factors are updated at
119 each iteration, however, this method is computationally rather cumbersome
120 and does not necessarily assure the global best design.

121 An OED problem including multiple experimental factors normally con-
122 tains a large number of design variables and has multiple local maxima/minima
123 (Banga et al., 2002), for which the commonly used gradient-based optimiza-
124 tion methods may only converge to local optima. Various global optimiza-
125 tion techniques have been developed to solve complex OED problems with
126 the purpose to obtain global optima and improve the convergent speed, see
127 (Banga et al., 2005; Catania and Paladino, 2009; Ruffio et al., 2012) for ex-
128 ample. Most global optimization techniques are population-based requiring a
129 large number of calculations of model equations and objective functions. The
130 computational load is increased exponentially with the increase in the num-
131 ber of design variables. This makes OED of multiple experimental factors
132 computationally demanding. In this work, we aim to develop a framework
133 to conduct OED of multiple experimental factors in an integrated, compu-

134 tationally efficient environment so that the data collected from the designed
135 experiments contain rich information for modeling. This framework will sup-
136 port modeling related tasks such as simulation of complex dynamic systems,
137 fundamental system analysis that are crucial for OED and parameter estima-
138 tion, e.g. parametric sensitivity analysis, structural identifiability analysis,
139 parameter estimability analysis, OED of multiple input factors and observa-
140 tion factors, assessment of OED results, etc.

141 The remaining of the paper is organized as follows. Section 2 presents
142 preliminaries on least-square parameter estimation and OED relevant analy-
143 sis such as parametric sensitivity analysis, structural identifiability analysis
144 and parameter estimability analysis. In Section 3, development of several key
145 OED problems are presented on single experimental factors including input
146 intensities, measurement set selection and sampling time profile, individu-
147 ally, using different optimization strategies. A novel integrated observation
148 design is proposed in Section 4, where the measurement set selection and
149 the sampling profile are determined simultaneously. In Section 5, an itera-
150 tive two-layer design is proposed for integrated design of input factors and
151 observation factors. OED on two case study systems, an enzyme reaction
152 system and a lab-scale enzymatic biodiesel production system, are simulated
153 and discussed in Section 6. Finally, conclusions and discussions are made in
154 Section 7. Details of case study models are given in Appendix.

155 2. Preliminaries on relevant methods

156 Consider a general nonlinear dynamic model with n state variables, p
157 parameters and m output variables, the state and output can be described

158 by a set of ordinary differential equations (ODEs) and algebraic equations:

$$\dot{\mathbf{X}}(t) = \mathbf{f}(\mathbf{X}(t), \boldsymbol{\theta}), \mathbf{X}(t_0) = \mathbf{X}_0, \quad (1)$$

$$\mathbf{Y}(t) = \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}) + \boldsymbol{\xi}(t). \quad (2)$$

159 where $\mathbf{f}(\cdot)$ is a set of state transition functions of the system dynamics which
 160 are assumed to be continuous and first-order derivative; $\mathbf{X} = [x_1, x_2, \dots, x_n]^T$
 161 $\in \mathbb{R}^n$ denotes the vector of n state variables with initial condition \mathbf{X}_0 ;
 162 $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_p]^T \in \mathbb{R}^p$ is the vector of p model parameters; $\mathbf{Y} \in \mathbb{R}^m$
 163 is the measurement output vector with $m(m \leq n)$ measurement variables;
 164 $\mathbf{h}(\cdot)$ is the measurement function, normally used for selecting which vari-
 165 ables to be measured. $\boldsymbol{\xi}$ is the vector of measurement errors which can be
 166 classified into systematic errors and random errors. The experiments should
 167 be designed to eliminate the systematic errors. However, the random errors
 168 that contaminate the observations always exist. Most often the measurement
 169 error is assumed to be a zero mean, Gaussian noise.

170 2.1. Least-square parameter estimation

171 Model parameters can be estimated using collected measurement data.
 172 When the system model is linear in parameters or can be transformed to be
 173 linear in parameters, a widely used method for parameter estimation is the
 174 (weighted) least-square estimation, where the problem is formulated as

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg \min_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) \\ &= \arg \min_{\boldsymbol{\theta} \in \Theta} \sum_{l=1}^N \left(\mathbf{Y}(t_l) - \hat{\mathbf{Y}}(\hat{\boldsymbol{\theta}}, t_l) \right)^T \cdot \mathbf{Q}^{-1} \cdot \left(\mathbf{Y}(t_l) - \hat{\mathbf{Y}}(\hat{\boldsymbol{\theta}}, t_l) \right), \quad (3) \end{aligned}$$

175 where \mathbf{Y} and $\hat{\mathbf{Y}}$ are measured values and model prediction of the output vec-
 176 tor at sampling times t_l ($l = 1, 2, \dots, N$), N is the total number of sampling

177 data in time. Assuming all observation variables can be measured inde-
 178 pendently and characterized by the variance of σ_j^2 , the measurement error
 179 covariance matrix is written as $\mathbf{Q} = \text{diag}[\sigma_1^2, \dots, \sigma_m^2]$.

180 The adequacy of the model and the parameter significance can be assessed
 181 by evaluating the output residuals through statistical tests. The method
 182 based on joint confidence regions between parameters is widely used to eval-
 183 uate the estimation quality (Franceschini and Macchietto, 2008). The confi-
 184 dence region can be determined based on the following cost function:

$$\left\{ \boldsymbol{\theta} : J(\boldsymbol{\theta}) \leq \left(1 + \frac{p}{N-p} F_{p, N-p}^{1-\alpha} \right) \times J(\hat{\boldsymbol{\theta}}) \right\}, \quad (4)$$

185 where $F_{p, N-p}^{1-\alpha}$ is the upper α -critical level of F distribution with p and $(N-p)$
 186 degrees of freedom; α is a positive real number between 0 and 1. However,
 187 for a nonlinear model, $J(\boldsymbol{\theta})$ is not a quadratic function with respect to $\boldsymbol{\theta}$,
 188 a linearization approximation is made by Taylor expansion around the es-
 189 timated parameters $\hat{\boldsymbol{\theta}}$. The confidence region can then be approximated as
 190 (Ljung, 1987)

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \cdot \mathbf{V}^{-1}(\hat{\boldsymbol{\theta}}) \cdot (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq p \times F_{p, N-p}^{1-\alpha} \quad (5)$$

191 where

$$\mathbf{V} = 2 \times \frac{J(\hat{\boldsymbol{\theta}})}{N-p} \times \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}, \quad \mathbf{H}(\hat{\boldsymbol{\theta}}) = \frac{\partial^2 J}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}^T} \quad (6)$$

192 Here \mathbf{V} is the parameter estimation error covariance matrix which is used as
 193 the cornerstone to measure parameter estimation uncertainty. $J(\hat{\boldsymbol{\theta}})/(N-p)$
 194 is an approximation of residual variance. \mathbf{H} is the Hessian matrix. The
 195 confidence interval of a single parameter θ_i can be determined by

$$\delta_i = \pm t_{N-p}^\alpha \times \sqrt{\mathbf{V}_{ii}} \quad (7)$$

196 where t_{N-p}^α is the student distribution with $(1-\alpha)$ confidence level and
 197 $(N-p)$ degrees of freedom. In later discussions, the formulation in (4)-(7)

will be used to produce confidence intervals to assess uncertainty in parameter estimation.

2.2. Structural identifiability analysis

As a key step and normally the initial step in parameter estimation scheme, structural identifiability analysis is performed to figure out whether it is possible to obtain unique parameter values for the candidate model structure from the data. If the parameters can be uniquely estimated from noise-free experimental data, then the model is said to be structurally identifiable. Consider the general dynamic model in (1) - (2), if

$$\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p, \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_1) = \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_2) \Leftrightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2, \quad (8)$$

the parameters in $\boldsymbol{\theta}$ are said to be globally identifiable. If the condition holds only for a neighbourhood of $\boldsymbol{\theta}^*$ in the parameter space which is given by

$$\begin{aligned} \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \{\boldsymbol{\theta} \in \mathbb{R}^p \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta\}, \\ \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_1) = \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_2) \Leftrightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2, \end{aligned} \quad (9)$$

the parameters $\boldsymbol{\theta}$ are said to be locally identifiable (McLean and McAuley, 2012). A number of methods have been developed to check the structural identifiability of nonlinear models such as Taylor series expansion approach (Pohjanpalo, 1978), generating series method (Walter and Lecourtier, 1982), local state isomorphism (Vajda et al., 1989) differential algebra algorithm (Ljung and Glad, 1994), or check the structural identifiability of the linearized part of the nonlinear model (Ben-Zvi et al., 2006). With the development of symbolic computational tools, the power series expansion methods that include the Taylor series expansion approach and the generating series method have been developed for structural identifiability analysis.

219 The basic idea of the Taylor series expansion approach is that the observa-
 220 tions of the system under consideration have unique analytic representations
 221 with respect to time, and therefore their derivatives with time are also rep-
 222 resented uniquely. Thus it is possible to represent the observations by using
 223 Maclaurin series expansion, written as

$$y_i(\boldsymbol{\theta}, t_0 + \Delta t) = y_i(\boldsymbol{\theta}, t_0) + \frac{dy_i}{dt} \Delta t + \frac{1}{2} \frac{d^2 y_i}{dt^2} (\Delta t)^2 + \dots \quad (10)$$

224 where Δt is a small time increment. The uniqueness of those Taylor series
 225 coefficients in (10) can guarantee the structural identifiability of the model.

226 With the generating series approach, the observations are expanded with
 227 respect to time and inputs. This method is refined to state models which are
 228 linear in the inputs, given as follows:

$$\dot{\mathbf{X}}(t) = \mathbf{f}(\mathbf{X}(t), \boldsymbol{\theta}) + \sum_{i=1}^{n_u} g_i(\mathbf{X}(t), \boldsymbol{\theta}) u_i(t) \quad (11)$$

229 where u_i stands for input factors, n_u is the number of input factors, and g_i is
 230 the corresponding coefficient for u_i . The observations in (2) can be expanded
 231 in such a way that the series coefficients are $\mathbf{h}(\mathbf{X}_0, \boldsymbol{\theta})$ and its Lie derivatives,
 232 $\mathbf{L}_{f_{j0}} \mathbf{h}, \mathbf{L}_{f_{j1}} \mathbf{h}, \dots, \mathbf{L}_{f_{jk}} \mathbf{h}$, where $\mathbf{L}_{\mathbf{f}} \mathbf{h}(\mathbf{X}_0, \boldsymbol{\theta}) = \sum_{j=1}^{n_u} g_j(\mathbf{X}_0, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial x_j} \mathbf{h}(\mathbf{X}_0, \boldsymbol{\theta})$.
 233 Similar to the Taylor series approach, the structural identifiability prob-
 234 lem is transformed into the determination of power series coefficients, the
 235 unique value of which provides a sufficient condition of structurally identifi-
 236 able model. However, it should be noted that there is no upper bound for
 237 the number of derivatives that needs to be calculated for nonlinear models.
 238 For nonlinear systems with a large number of parameters the calculation of
 239 power series coefficient is a computational cumbersome work.

240 In this work, the identifiability tableau method proposed in (Balsa-Canto
 241 et al., 2010) is used for structural identifiability analysis. The identifiability

242 tableau is constructed to represent the non-zero elements of the Jacobian
 243 matrix of those power series coefficients on model parameters. Some model
 244 parameters can be obtained directly from solving simple algebraic equations.
 245 With the obtained model parameter values, the identifiability tableau can be
 246 reduced and eventually minimized. The analysis of the remaining parameters
 247 will be conducted in a sequential procedure. The structural identifiability of a
 248 model parameter depends on the existence of the solution of that parameter.
 249 More details on identifiability tableau can be found in (Balsa-Canto et al.,
 250 2010; Chis et al., 2011).

251 *2.3. Parameter estimability analysis*

252 For a structurally identifiable model, its unknown parameters may still
 253 not be estimable in practice (also called practical identifiability) due to sev-
 254 eral reasons: (i) the experimental data for parameter estimation are sparse
 255 and noisy, or contains inadequate information due to poorly designed exper-
 256 iments; (ii) some unknown parameters have very little influence on model
 257 outputs, i.e., of low parametric sensitivities; (iii) the effect of some param-
 258 eters to the model prediction can be compensated by other parameters, i.e.,
 259 high correlations exist between parameters to be estimated.

260 Practical identifiability analysis is in general a discrete (combinatorial)
 261 non-convex optimization problem. Exhaustive search and genetic algorithms
 262 are the most widely used methods to get the solution. However, for nonlinear
 263 dynamic systems with a large number of parameters, these methods are com-
 264 putationally too expensive. Methods of approximations and relaxations of
 265 the original optimization problem have been developed and applied to evalu-
 266 ate practical identifiability. These include but not limited to the collinearity

index (Brun et al., 2001), the relative gain array (Sandink et al., 2001), the Hanken singular value (Sun and Hahn, 2006), orthogonalization based methods (Yao et al., 2003), optimization methods that rely on the Fisher information matrix, and methods with repeated parameter estimation. In this work, an orthogonalization based method (Yao et al., 2003) will be used for practical identifiability analysis. This method is based on the measure of orthogonal parameter sensitivities. In another word, the parameter pair correlations have been removed from the original local sensitivity matrix and the measurement is focused on the independent parameters.

2.4. Parametric sensitivity analysis

Parameter sensitivity analysis is a method used to examine how sensitive the system output is in response to variations in model parameters. The parametric local sensitivities can be described by

$$\dot{\mathbf{S}} = \frac{\partial \mathbf{f}}{\partial \mathbf{X}} \cdot \frac{\partial \mathbf{X}}{\partial \boldsymbol{\theta}} + \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} \quad (12)$$

where $\mathbf{S} = \frac{\partial \mathbf{X}}{\partial \boldsymbol{\theta}} = [s_{ij}] \in \mathbb{R}^{n \times p}$ is the parameter local sensitivity matrix, $s_{ij} = \frac{\partial x_i}{\partial \theta_j}$; $\frac{\partial \mathbf{f}}{\partial \mathbf{X}} \in \mathbb{R}^{n \times n}$ is the Jacobian matrix, and $\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{n \times p}$ is the parametric Jacobian matrix. The state differential equations in (1) and the sensitivity differential equations in (12) can be solved simultaneously through the direct differential method (Atherton et al., 1975). To remove the effects of model parameters that are likely to have values at different scales, normalized sensitivities, $\bar{s}_{ij} = \frac{\partial x_i}{\partial \theta_j} \cdot \frac{\theta_j}{x_i}$, are sometimes used for comparison of parameter sensitivities. The corresponding normalized sensitivity matrix is $\bar{\mathbf{S}} = [\bar{s}_{ij}]_{n \times p}$. The overall effect of parameter θ_j to all state variables can be calculated by a norm of local sensitivities such as

$$OS_j = \frac{1}{N} \sqrt{\sum_{i=1}^n \sum_{l=1}^N s_{ij}^2(t_l)} \quad (13)$$

290 The sensitivity analysis results can be used to find key parameters that
 291 have significant impacts to system behavior, to assist model simplification
 292 or used in gradient based optimization process for parameter estimation. In
 293 model based OED, the parametric sensitivity matrix is taken to construct the
 294 FIM. Therefore, parameter sensitivity plays an indispensable role in param-
 295 eter estimation, parameter identifiability analysis and experimental design.

296 **3. OED for single experimental factors**

297 *3.1. Fisher information matrix and design criteria*

298 The task of model-based OED for parameter estimation is to determine
 299 the values of experimental variables so that the predicted measurement data
 300 information is optimized. Denoting the design factors which characterize the
 301 experiment into a vector $\boldsymbol{\zeta}$, the FIM can be locally written as

$$\mathbf{FIM}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \mathbf{S}(\boldsymbol{\theta}, \boldsymbol{\zeta})^T \cdot \mathbf{W} \cdot \mathbf{S}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \quad (14)$$

302 where the weighting matrix \mathbf{W} is normally taken to be \mathbf{Q}^{-1} for the most gen-
 303 eral discussion. The FIM can be used to quantify the information content
 304 of an experiment towards parameters to be estimated. The more sensitive
 305 of a state variable to a parameter, the more information is contained in the
 306 FIM about that parameter. The inverse of the measurement error covariance
 307 matrix, \mathbf{Q}^{-1} , in the FIM indicates that data with a larger measurement error
 308 will contribute less reliable information than the data with a smaller mea-
 309 surement error. In addition, the correlations between measurements are also

310 considered in the FIM. When the model is linear in its parameters, according
 311 to the Cramer-Rao lower bound inequality, the FIM is approximately equal
 312 to the inverse of the parameter estimation error covariance matrix, Σ , under
 313 the assumption of unbiased parameter estimation and uncorrelated additive
 314 white measurement noise (Ljung, 1987).

315 The OED problem can be cast as minimization of a proper measure of
 316 the parameter error covariance matrix, which can be approximated as the
 317 inverse of FIM, i.e.

$$\zeta^* = \arg \min_{\zeta \in \Omega} \Phi \left((\mathbf{FIM}(\boldsymbol{\theta}, \zeta))^{-1} \right), \quad (15)$$

318 where Ω is the admissible space of the design factors, $\Phi(\cdot)$ represents a func-
 319 tion to scalarize the inverse of FIM. The most commonly used design criteria
 320 in OED are A-optimal, D-optimal, E-optimal, and modified E-optimal de-
 321 signs, in which the scalar measures are closely related to the shape, size and
 322 orientation of parameter estimation confidence intervals. The design focus of
 323 these scalar design criteria are different from each other due to the different
 324 features taken from the FIM. No single design criterion can be applicable to
 325 all design problems or suitable for all systems. For a given dynamic system,
 326 one particular optimization criterion may be superior to others; but this does
 327 not necessarily mean that this criterion plays well in other designs. There-
 328 fore, it is recommended that different criteria should be tried and compared
 329 in a standard experimental design.

330 For most chemical and biochemical reaction systems, the OED for pa-
 331 rameter estimation can be put into two categories, i.e. input design on
 332 manipulation of input variables, and observation design such as design of
 333 sampling time profile and selection of measurement variables. For a given

dynamic system, the change in input will change the dynamic response. This means during the OED process, for each value taken for an input factor, the full dynamic response profile needs to be calculated. On the other hand, in the design of observation variables, the dynamic response is determined by the specific input condition, only one calculation of the dynamic response is required for the optimization process. For this reason, the experimental design formulation and the optimization processes for the design of input factors and for the design of observation factors can be quite different.

3.2. *Input intensity design*

The purpose of OED of input factors is to choose the type and duration of input stimulation/perturbations. Inputs can be fixed or time-dependent for a chemical reaction system and many other dynamic systems. When the input design factor is time-dependent, a typical option is to transfer the original OED problem into a relaxed finite dimensional nonlinear programming dynamic optimization problem by approximating the time-varying inputs with discrete form of inputs. The problem can then be solved by direct dynamic optimization methods such as the sequential methods and the simultaneous methods (Biegler et al., 2002).

In this work, the input factors considered for chemical reaction systems are those initial conditions of the reaction species that can be manipulated through experimental setting. The OED problem is formulated as the general form in (15), in which the design factors are the initial input intensities, i.e., $\boldsymbol{\zeta} = \mathbf{X}_0$. It should be noted that only those elements in \mathbf{X}_0 that need to be designed are included in the OED, other initial conditions are kept at the values according to the system mechanism and operating conditions.

359 Since the response of a dynamic system will change following the change in
360 inputs, the FIM is also changed and needs to be calculated for each choice
361 of the input. Numerically this will involve integration of ODEs in (1) being
362 implemented many times during the optimization process.

363 This input design is in general an non-convex optimization problem that
364 is difficult to solve to get the global solution. To obtain the optimal initial
365 conditions of multiple inputs, in this work the particle swarm optimization
366 (PSO) algorithm is chosen, which has not been used in previous multi-input
367 OED.

368 3.3. *Measurement set selection*

369 Collecting measurement data with rich information for modeling could
370 be cost expensive and time-consuming, especially for complex biological or
371 biochemical systems. The aim of OED on measurement set selection is to
372 find a necessary or a minimum set of variables to be measured such that
373 the selected measurement variables are most useful or discriminating for pa-
374 rameter estimation. From the system development point of view, another
375 benefit from optimized measurement set selection is that the design results
376 may indicate missing measurement of variables that are actually crucial to
377 modeling. Necessary measurement can then be added to the sensing system.
378 In this work, it is assumed that each state variable can be measured inde-
379 pendently. For some circumstances where only combination of states can
380 be measured, similar design can still be applied since the importance of the
381 combined measurement of interest can be easily determined from the ranking
382 (and the weighting) of each individual state after the OED.

383 Assuming that the measurement set is selected from the full set of the

state variables, the measurement set selection problem can be formulated as follows (Flaherty et al., 2006):

$$\begin{aligned} \xi &= \begin{Bmatrix} x_1 & \cdots & x_n \\ \lambda_1 & \cdots & \lambda_n \end{Bmatrix} \\ \xi^* &= \arg \min_{\lambda \in \Omega} \Phi \left(\left(\sum_{i=1}^n \frac{1}{\sigma_i^2} \lambda_i \mathbf{S}_i^T \mathbf{S}_i \right)^{-1} \right) \\ s.t. \quad & \lambda_i \in \{0, 1\}, \quad \mathbf{1}^T \boldsymbol{\lambda} = n_{sel} \end{aligned} \quad (16)$$

where $\mathbf{1}$ is a column vector comprised of ones in all its entries; $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]^T$, in which λ_i is the non-negative weight factor for x_i that can be chosen as either 1 or 0; n_{sel} is the total number of measurement variables to be used. After the OED, those state variables with weighting factor values to be 1 are selected to form the measurement set.

3.4. Sampling time profile design

The target of optimal design of sampling time profile(s) is to determine the sampling time points that will enable most informative data collection at those points. The design problem can be set up as to choose certain number of sampling points along the measurement states, which, in principle, is an infinite dimensional non-convex dynamic optimization problem hard to solve. To tackle this difficulty, the sampling time profile design can instead be formulated as a discrete optimization problem. The available measurement variables are defined *a priori*, also the total number of sampling points is given for each measurement variable, and the OED is performed to find the best combination of a subset of the data points from the whole set.

Similar to the OED of measurement set selection, the optimal design problem of sampling time profile can be formulated as follows

$$\begin{aligned} \xi &= \begin{Bmatrix} t_1 & \cdots & t_N \\ \omega_1 & \cdots & \omega_N \end{Bmatrix} \\ \xi^* &= \arg \min_{\omega \in \Omega} \Phi \left(\left(\sum_{i=1}^N \frac{1}{\sigma_i^2} \omega_i \mathbf{S}(t_i)^T \mathbf{S}(t_i) \right)^{-1} \right) \\ s.t. \quad & \omega_i \in \{0, 1\}, \mathbf{1}^T \omega = N_{sp} \end{aligned} \quad (17)$$

where $\omega = [\omega_1, \dots, \omega_N]^T$ is the weighting vector for all the available measurement points in time horizon. $N_{sp} (\leq N)$ is the total number of sampling points to be selected. Here it is assumed that the same sampling time profile is applied to all considered measurement variables. One should note that time resolution should be small enough so that the optimal sampling time solution are included in the predefined sampling time set.

4. Integrated observation design

We start from OED of observations by fixing the input experimental factors. The observation design of measurement strategies include but are not limited to the measurement set selection and the sampling time profile design. Compared with the input experimental design, one big advantage in design of measurement factors is that when the input (stimulation/perturbation) is fixed, the dynamic response of the system is also determined, in other words, the candidate pool of the available measurement information is provided. The observation design is mainly to find a strategy that can pick up the most informative data from the available measurement data.

As can be seen from Sections 3.3 and 3.4, the design of sampling time profile and the design on measurement set to be selected are handled separately. In each design, it is assumed that all the other experimental factors are specified. This single factor design may fail to give a satisfactory result to guide measurement data collection since the experimental factors in observation could be correlated to each other in terms of providing information content. A more effective OED should put the multiple observation factors together into one integrated design. One option is to go through an iterative procedure to design the two experimental factors, in each iteration only one factor is optimized based on the predefined settings of the other one, and repeats until both factors are properly designed. This iterative procedure is not computationally efficient, also the dependent effects of the two measurement factors are still handled separately during the design.

Here we propose to combine the measurement set selection and the sampling time profile design into one single optimization problem. This idea is inspired by the fact that the two optimization problems share a similar formulation as in (16) and (17), and only one integration of the state variables is required during the optimization design under given input. The integrated observation design is represented as the following optimization problem.

$$\begin{aligned} \xi &= \begin{Bmatrix} t_1 & \cdots & t_{N \times n} \\ \omega_1 & \cdots & \omega_{N \times n} \end{Bmatrix} \\ \xi^* &= \arg \min_{\omega \in \Omega} \Phi \left(\left(\sum_{i=1}^{N \times n} \frac{1}{\sigma_i^2} \omega_i \mathbf{S}(t_i)^T \mathbf{S}(t_i) \right)^{-1} \right) \\ s.t. \quad & \omega_i \in \{0, 1\}, \quad \mathbf{1}^T \boldsymbol{\omega} = N_{ssp} \end{aligned} \quad (18)$$

Here the number of the integrated weighting factors is extended to $n \times N$

440 for the system with n state variables and the data length of N , i.e., $\boldsymbol{\omega} =$
 441 $[\omega_1, \omega_2, \dots, \omega_{n \times N}]^T$. Each ω_i stands for the importance of one measurable
 442 state variable at a particular time point. $N_{ssp} (\leq n \times N)$ is the total number
 443 of sampling points to be selected for all the state variables at all chosen time
 444 points. The design problem as formulated in (18) is an integer programming
 445 problem which can be solved by exhaustive search if the number of $n \times N$
 446 is relatively small. For a design that contains a large number of weighting
 447 factors, the optimization problem in (18) can be further relaxed to an approx-
 448 imate continuous optimization problem (Yue et al., 2008; He et al., 2010),
 449 which is given as follows.

$$\begin{aligned}
 \boldsymbol{\xi} &= \begin{Bmatrix} t_1 & \cdots & t_{N \times n} \\ \omega_1 & \cdots & \omega_{N \times n} \end{Bmatrix} \\
 \boldsymbol{\xi}^* &= \arg \min_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \Phi \left(\left(\sum_{i=1}^{N \times n} \frac{1}{\sigma_i^2} \omega_i \mathbf{S}(t_i)^T \mathbf{S}(t_i) \right)^{-1} \right) \\
 s.t. \quad & \sum_{i=1}^{N \times n} \omega_i = 1, \omega_i \geq 0
 \end{aligned} \tag{19}$$

450 The weighting term ω_i is relaxed to a continuous variable taking values be-
 451 tween $[0, 1]$. In this way, the the optimal solution provides a lower bound
 452 for the original integer optimization problem. At each sampling time point,
 453 the FIM for involved state variables is a positive definite matrix. Therefore,
 454 the continuous optimization problem in (19) can be converted into a con-
 455 vex optimization problem by employing different scalar design criteria. For
 456 instance, taking the D-optimal design criterion, the observation design prob-
 457 lem can be easily transformed into a convex optimization problem that can
 458 be solved by local optimization methods such as the Powell's quadratically

459 convergent method (Kutalik et al., 2004) or the interior-point method. When
 460 the A-optimal or E-optimal design criterion is applied, problem (19) can be
 461 transferred into an equivalent semi-definite programming (SDP) problem.
 462 The E-optimal observation design formulation is written as follows.

$$\begin{aligned}
 \min \quad & -t \\
 \text{s.t.} \quad & \sum_{i=1}^{n \times N} \frac{1}{\sigma_i^2} \omega_i \mathbf{S}_i^T \mathbf{S}_i \succ t \mathbf{I} \\
 & \omega_i \succ 0, \forall i; \quad \mathbf{1}^T \boldsymbol{\omega} = 1
 \end{aligned} \tag{20}$$

463 The optimization problem in (20) can be conveniently solved by available
 464 computational tools such as the 'SeDuMi' software. When the gradient-
 465 based optimization method is used to solve the problem, the derivative of
 466 the objective function over the weights is much easier to calculate than the
 467 direct derivative over time and state variables. With this integrated design,
 468 the sampling time profile and the measurement set are simultaneously deter-
 469 mined through a single-objective optimization.

470 **5. Iterative double-layer design of both observation and input**

471 In a systematic experimental design, those major experimental conditions
 472 such as the input perturbations and the measurement strategy should be
 473 considered in an integrated design framework. This integrated optimization
 474 problem can be handled through a sequential process where the input design
 475 and the observation design are solved sequentially and iteratively until the
 476 satisfactory result is obtained. The input design problem can be formulated
 477 as a complex non-convex optimization problem as discussed in Section 3.2,

478 while the measurement design problems are treated as a convex optimization
 479 problem as described in Sections 3.3 and 3.4, separately, or with the simul-
 480 taneous design as proposed in Section 4. As such, there is no simple solution
 481 for this multi-factor optimization problem.

482 In this work, we propose an iterative double-layer procedure, as illus-
 483 trated in Fig.2, to design the experimental factors for both the input and the
 484 observation. The design of input factors is processed in the upper layer, and
 485 the integrated observation design is handled in the lower layer.

486 Due to the non-convex nature of the input design problem, a modern
 487 heuristic method - PSO (Kennedy, 2011), is chosen to obtain the optimal
 488 solution globally. The PSO method is a population-based optimization al-
 489 gorithm which can solve a variety of hard problems with fast convergent
 490 rates. With this algorithm, only a few parameters need to be tuned and no
 491 derivative calculations are required, making the algorithm attractive from
 492 the computation point of view. The basic PSO method is based on a pop-
 493 ulation of s particles that represent solutions of the optimization problem.
 494 Each particle is associated with a position x and a velocity v , which denote
 495 its position and movement through the searching space. The position and
 496 velocity of a particle can be dynamically adjusted via an iterative process
 497 according to the objective function values at particle positions. At the gen-
 498 eration k , the new position x_i^{k+1} of the i -th particle is computed by adding
 499 to the old position x_i^k a velocity vector v_i^{k+1} :

$$x_i^{k+1} = x_i^k + v_i^{k+1} \quad (21)$$

500 The velocity vector of the i -th particle is updated by

$$v_i^{k+1} = \omega \cdot v_i^k + \alpha_1 \cdot r_1 \cdot (pbest_i^k - x_i^k) + \alpha_2 \cdot r_2 \cdot (gbest^k - x_i^k) \quad (22)$$

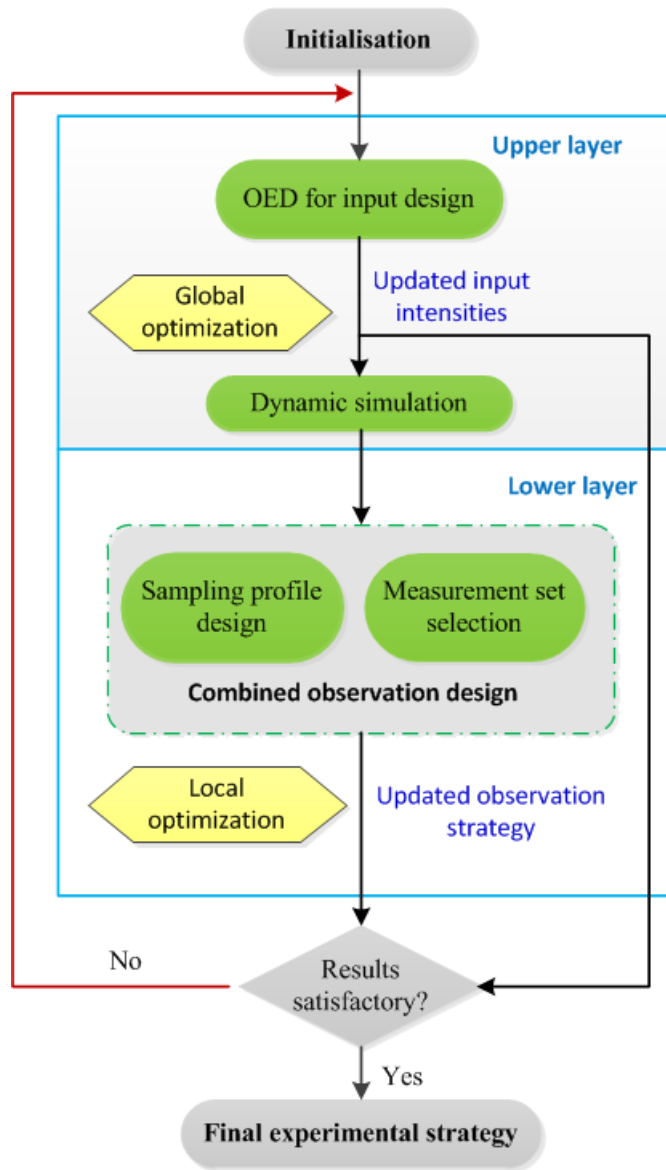


Figure 2: Iterative double-layer design for both input factors and observation factors

where ω , α_1 and α_2 are the inertia parameter, the cognition parameter and the social parameter, respectively. r_1 and r_2 are numbers randomly chosen in the range of 0 to 1. $pbest_i^k$ is the best position of the i -th particle at the k -th generation, and $gbest^k$ is the best position of the k -th generation among all particles, which can be determined by

$$gbest^k = \arg \min_{z \in x_1^k, x_2^k, \dots, x_s^k} g(z) \quad (23)$$

Here $g(\cdot)$ is the objective function. The pseudo code for PSO implementation is described in Algorithm 5.1 as follows.

Algorithm 5.1

1. Choose a population size s and the iteration number n_{tol} . Initialize the swarm positions $x_1^0, x_2^0, \dots, x_s^0$ and their velocities $v_1^0, v_2^0, \dots, v_s^0$.
2. Let $pbest_i^0 = x_i^0, i = 1, 2, \dots$, determine $gbest^0$ using (23), and let $k = 0$.
3. Set $gbest^{k+1} = gbest^k$. For every particle i , do:
 - Check the constraint of x_i^k , make sure that each particle stays within the bound.
 - If $f(x_i^k) \leq f(pbest_i^k)$, then update the best position of the i -th particle, $pbest_i^{k+1} = x_i^k$; if $f(pbest_i^{k+1}) \leq f(gbest^{k+1})$, then update the best position at current generation, $gbest^{k+1} = pbest_i^{k+1}$; otherwise, set $pbest_i^{k+1} = pbest_i^k$.
4. Compute x_i^{k+1} and v_i^{k+1} for each particle using equations (21) and (22).
5. Terminate the process when $k = n_{tol}$. Otherwise, increase k by one and go to step 3.

With this iterative double-layer structure, the inputs are firstly determined by applying the PSO for a pre-defined number of iterations, based on

524 which the observation design problem is solved at the lower layer through
 525 the Powell's conjugate direction method (Fletcher and Powell, 1963). The
 526 designed observation strategy is then used in the next iteration for an up-
 527 dated design of the input factor. This process lasts until the optimal solution
 528 is obtained. While the optimization at the lower layer can solve the convex
 529 optimization problem of observation design under the given input conditions,
 530 the upper-layer design employing stochastic searching largely increases the
 531 chance of finding a global solution for input factors. This is a clear advan-
 532 tage over the traditional local numerical algorithms which most likely only
 533 lead to local optimum. For a complex OED problem including both input
 534 design and observation design, it is also computationally more efficient to put
 535 the observation design at the lower layer since this is a convex optimization
 536 problem that is relatively easy to solve. The main procedure of the iterative
 537 double-layer optimization design is given in the following.

538 **Algorithm 5.2**

- 539 1. Initialize the overall OED objective function $g(x, y)$, where x and y de-
 540 note the input and observation variables, respectively. Set the stopping
 541 tolerance level $\delta_{tol} \geq 0$.
- 542 2. Let the iteration number $l = 0$, use the Powell's method to calculate
 543 y_{best}^0 based on x_{set} . x_{set} is a vector of pre-setting values for the input
 544 variables. Then determine x_{best}^0 for $g(x, y_{best}^0)$ using Algorithm 5.1.
- 545 3. For iteration l , determine x_{best}^l for the objective function $g(x, y_{best}^{l-1})$ in
 546 the upper layer using the PSO method described in Algorithm 5.1,
 547 then calculate y_{best}^l for $g(x_{best}^l, y)$ in the lower layer using the Powell's
 548 method.

549 4. If $|g^{l+1} - g^l| \leq \delta_{tol}$, then stop the optimization process. Otherwise,
 550 increase l by one and go back to step 3.

551 Using this iterative double-layer strategy, the input design and the obser-
 552 vation design problems can be integrated into one optimization framework.
 553 Different from the sequential design process where each OED problem is op-
 554 timized only once, the proposed method enables the update of the input
 555 variables and the observation strategies during each iteration of the opti-
 556 mization process. In this way, the design order of multiple factors does not
 557 need to be considered.

558 6. Case studies on two biochemical reaction processes

559 6.1. Enzymatic process with kinetically controlled synthesis reactions

560 The first case study system is an enzymatic process with kinetically con-
 561 trolled synthesis reactions as illustrated in Fig. 3. In this reaction system, **S**
 562 is the donor substrate, **P** is the leaving group product, **N** denotes the nucle-
 563 ophile, **Q** is the desired product, **R** is the hydrolysis by-product; **W** stands
 564 for water whose quantity is taken as constant due to its large amount; **E** is the
 565 enzyme and **ES**, **E***, **EQ** and **ER** are different complex forms of enzymes.
 566 All reactants are assumed to be well mixed in the reactor. At the begin-
 567 ning of the reactions, the initial reactant species are the donor substrate, the
 568 nucleophile and the catalyst. The substrate firstly binds to the enzyme to
 569 form the enzyme-substrate complex, **ES**, and then **ES** can be decomposed
 570 into another compound **E*** and the leaving group product **P**. **E*** can either
 571 react with the nucleophile to form **EQ** or be hydrolyzed to produce **ER**. The
 572 compound **EQ** can be decomposed into the required product and enzyme,

573 while **ER** can be decomposed into the hydrolysis by-product and enzyme.
 574 During the whole reaction process, all the reactions are reversible, except
 575 for the decomposition of **ER** to give **E** and **R**. Due to the characteristics of
 576 enzyme, it only catalyzes the reactions and at the end of the reaction, the
 577 amount of enzyme remains the same as before the chemical reactions.

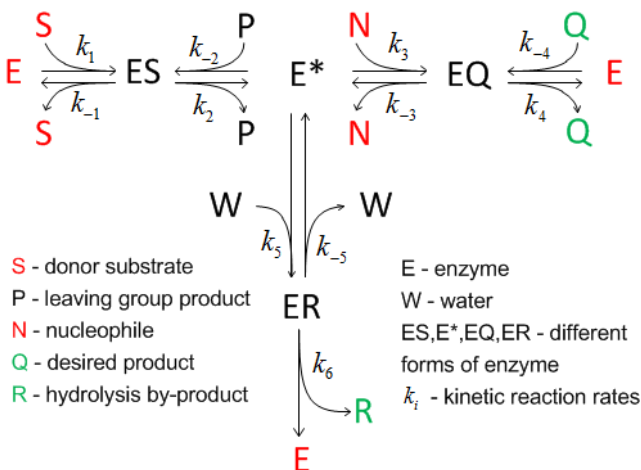


Figure 3: Enzyme kinetically controlled synthesis process

578 A number of enzymatic processes have the similar kinetically controlled
 579 synthesis reaction scheme, e.g., in the preparation of semi-synthesis peni-
 580 cillins, **S** is hydroxyphenylglycine methyl ester and **N** is 6-APA, etc. In this
 581 system, the desired product **Q** is not thermodynamically the most favourable
 582 one. The hydrolysis by-product **R** will dominate at long times. Among those
 583 reaction species, **Q**, **S**, **P**, **N** and **R** are measurable in experiments while it
 584 is difficult to measure different forms of enzymes due to its very low con-
 585 centrations. The initial concentrations of **S**, **N** and **E** are user-controllable
 586 inputs written as S_0 , N_0 and E_0 , respectively. It is known that a chem-

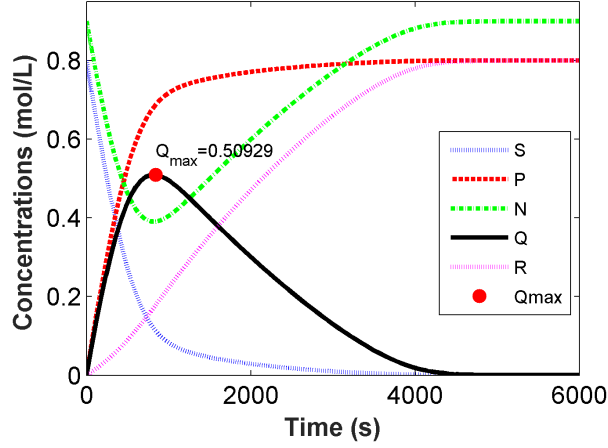


Figure 4: Time profiles of the 5 measurable state variables of the enzyme reaction system

ical reaction is always affected by the surrounding environment and some other factors such as the inactivation of enzyme, reactant instability, effects of pH and temperature, etc. In order to investigate the system with a focus on experimental design, all these complications have been removed in this test system. Following the mass balance principle, the enzyme reaction system can be expressed as 10 ODEs including 11 parameters, as given in Appendix B. The nominal values of the model parameters and the initial conditions of the state variables are listed in Table B.5. The time profiles of the 5 measurable state variables under the nominal model parameters and initial conditions are illustrated in Fig. 4. More modeling details and system analysis of this enzyme reaction system can be found in (Yue et al., 2013).

6.1.1. Structural identifiability analysis

To determine whether the model parameters are structurally identifiable, the generating series approach combined with the identifiability tableau, as

introduced in Section 2.2, are implemented to the enzyme reaction model. Ideally it is always possible to obtain a full rank Jacobian matrix for the power series coefficients because the number of Lie derivatives of model equations is infinite.

Through the numerical steps as proposed in (Balsa-Canto et al., 2010), the Jacobian matrix of the series coefficients with respect to model parameters can be obtained and shown in the tableaux shown in Fig. 5. Each row represents one series coefficient determined by the Lie derivative and each column represents one model parameter. In such a tableau, each black grid denotes that the series coefficient in that row contains non-zero element with respect to the model parameter in the corresponding column. In Fig. 5(a), there are 27 non-zero series coefficients with respect to the 11 model parameters, which are obtained by the Lie derivative computations. Fig. 5(b) shows a reduced tableau where 11 necessary rows are selected which can guarantee full rank of the Jacobian matrix. In this tableau a unique non-zero element in a given row means that the model parameter in the corresponding column can be identified, and this identifiable parameter can then be removed from the tableau. The elimination of a column (parameter) will lead to a reduced tableau with new unique non-zero elements. This process will continue, iteratively, until the tableau cannot be further reduced. For the enzyme reaction system, the final minimum tableau is shown in Fig. 5(c), in which only five parameters are remained that need to be further checked to see whether they are structurally identifiable or not. When all the five measurable state variables are included in the observation, all of the 11 model parameters can be determined as globally structurally identifiable. When a subset of the

626 five states are included in the observation, some parameters are found to be
 627 locally structurally identifiable or even not identifiable.

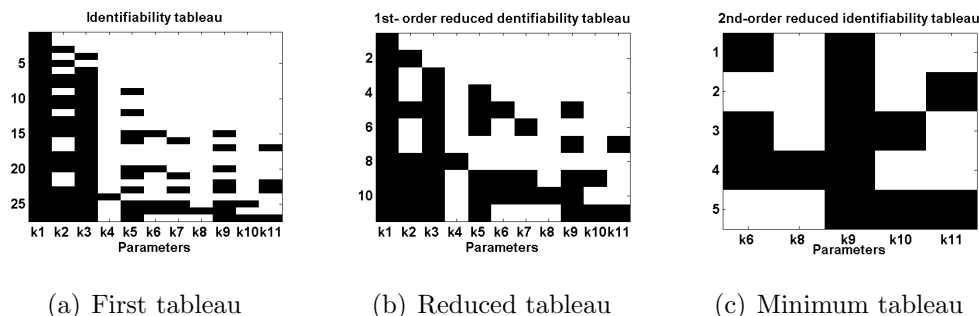


Figure 5: Identifiability tableaus based on generating series approach

6.1.2. Sensitivity analysis and parameter estimability analysis

629 Linear correlation between parameter pairs will affect parameter estima-
 630 bility. The correlation coefficient between two parameters, k_i and k_j , can be
 631 calculated from FIM as $R_{ij} = \frac{cov(k_i, k_j)}{\sqrt{FIM_{ii} \times FIM_{jj}}}$. Parameters k_i and k_j are
 632 said to be linearly correlated if $R_{ij} = 1$ or -1 . In the correlation matrix
 633 composed of R_{ij} , any non-diagonal entries with values close to ± 1 suggests a
 634 strong correlation between the pair of parameters. Simulation results show
 635 high correlations between several reversible reaction pairs in this system al-
 636 though they are not fully correlated. The orthogonalized sensitivity analysis
 637 (Yao et al., 2003) is implemented (see the algorithm in Appendix A) instead
 638 of the standard local sensitivity analysis (LSA). The ranking of parameters in
 639 terms of their influence to the states and the correlation coefficients between
 640 parameter pairs are shown in Fig. 6. The metric of 'IEOS' in Fig. 6 repre-
 641 sents the integrated effect of the model parameters to the model outputs by
 642 using the orthogonalized sensitivity analysis.

643 Considering both local sensitivities and correlations between all param-
644 eters, three parameters, k_2 , k_{-3} and k_5W are found to be the most important
645 and most identifiable among the 11 parameters. This result is largely consis-
646 tent with our previous analysis based purely on LSA, where k_2 , k_{-3} and k_{-5}
647 were identified to be the top three most important parameters (Yue et al.,
648 2013). With the orthogonalization-based method employed in this work, pa-
649 rameter k_5W replaces k_{-5} in the top 3 key parameters, which may due to the
650 fact that (k_5W, k_{-5}) is a highly correlated parameter pair. In the following
651 OED simulation studies on this case study enzyme reaction system, these
652 three key parameters are considered in the design.

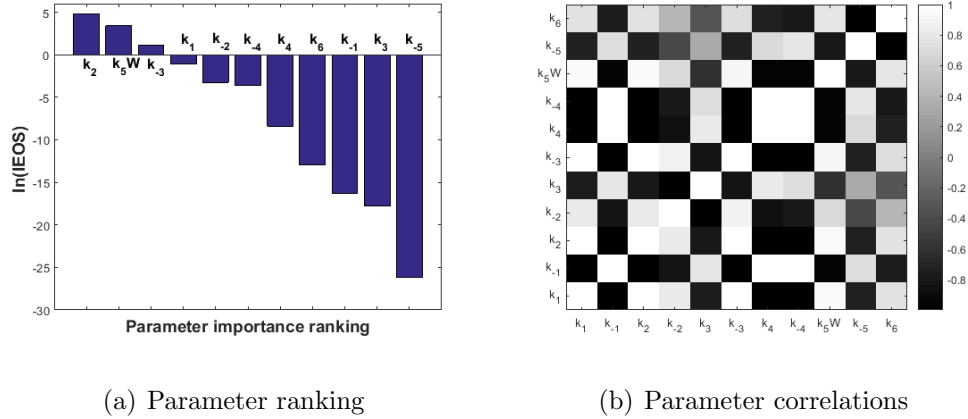


Figure 6: Orthogonalization-based sensitivity analysis & correlation analysis results

653 6.1.3. Observation design and results

654 We start from observation design by taking the nominal parameter val-
655 ues and the initial conditions in Appendix B. The design objectives are: (i)
656 to select measurement state variables; and (ii) to locate 100 measurement

657 data points for the selected state variables, which will lead to the most in-
658 formative data set for parameter estimation. In the simulation, 10% relative
659 measurement errors and 0.001 absolute measurement errors are added to the
660 simulation data. Three different design strategies, as shown in Table 1, are
661 compared during the simulation. A sequential design procedure is taken as
662 Strategy 1 and Strategy 2, where the former starts with the measurement set
663 selection followed by the sampling profile design, and the latter starts from
664 the sampling profile design followed by the measurement set selection. In
665 the proposed integrated design, named as Strategy 3, the design of the two
666 tasks are combined into one single optimization problem and the solutions
667 for both can be obtained simultaneously. The D-optimal design criterion is
668 employed in all the three OED methods.

Table 1: Three observation design strategies

OED methods	Design procedures
Strategy 1	Sequential: measurement set \longrightarrow sampling time profile
Strategy 2	Sequential: sampling time profile \longrightarrow measurement set
Strategy 3	Simultaneous: measurement set & sampling profile

669 The design results of the three observation strategies and also the default
670 setting without any OED are listed in Table 2. When no OED is employed,
671 all the 5 measurable states are taken into account, and the same uniform
672 sampling rule is applied to all the 5 states, i.e., 20 sampling points for each
673 state. For OED with Strategy 1, the two variables, S and Q, are firstly
674 selected to form the measurement set, then the sampling profile design is
675 performed to {S, Q}, which gives 3 sampling regions. In Strategy 2, the

sampling design is made first to all the 5 states and 3 sampling regions are found. Then using the designed sampling profile, the measurement set is selected which in fact includes two states, S and Q. Instead of taking these two variables, all five measurable variables are included otherwise the total number of data will be reduced. With the proposed Strategy 3, the total number of 100 sampling points are 'allocated' to S and Q after the optimal design. It can be seen that both Strategy 1 and Strategy 3 select S and Q as the most important measurement variables although the sampling profiles are different.

For Strategy 1, Strategy 2 and the no-OED cases, all (or selected) variables have the same sampling profile. Only with Strategy 3, the sampling profiles for each selected variable can be different. Taking S and Q as the state variables, the sampling points distribution from different experimental strategies are shown in Fig. 7. With the sequential design of Strategy 1 and Strategy 2, three sampling regions are recommended at different reaction stages, mostly corresponding to where the variables or local sensitivities have large changes. The sampling regions of Strategy 2 are narrower compared to Strategy 1. This is because there are five measurement variables in Strategy 2 and only two variables in Strategy 1 at the design of sampling profile. Using the proposed Strategy 3, two sampling regions are designed for S and two for Q, respectively, covering a wide range of the reaction process. Within each sampling region, consecutive measurement points are recommended by the design result which indicates that measurement within those selected sampling regions can potentially provide informative data.

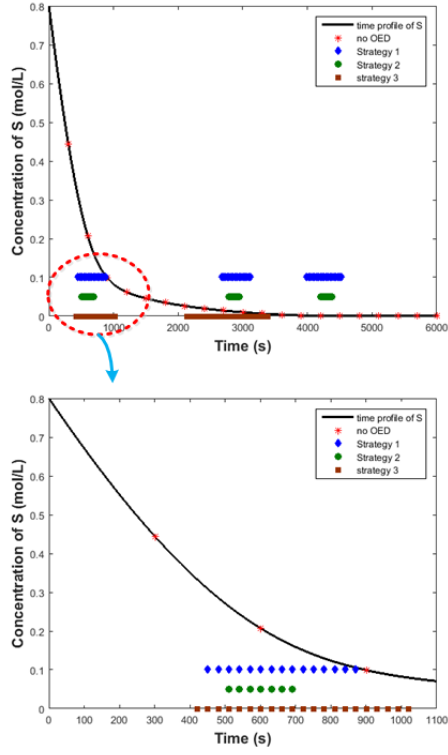
The confidence interval (CI) of the three key parameters in pairs are com-

701 pared in Fig. 8. According to the Cramer-Rao inequality, a smaller CI region
702 corresponds to smaller lower bounds for parameter estimation errors, there-
703 fore a better estimation quality can possibly be obtained. In this simulation,
704 the design Strategy 1 shows better result than Strategy 2, which suggests
705 that in the sequential design, the measurement set should be selected prior
706 to the sampling time design. The proposed integrative design, Strategy 3,
707 achieves the best result among the three methods due to the fact that all ob-
708 servation factors are considered simultaneously during the OED. The OED
709 of observation provides a useful insight that the sampling points should be
710 taken at certain regions that corresponds to large parameter sensitivities or
711 large change rates in key variables, not necessarily equally spaced as in a
712 traditional way.

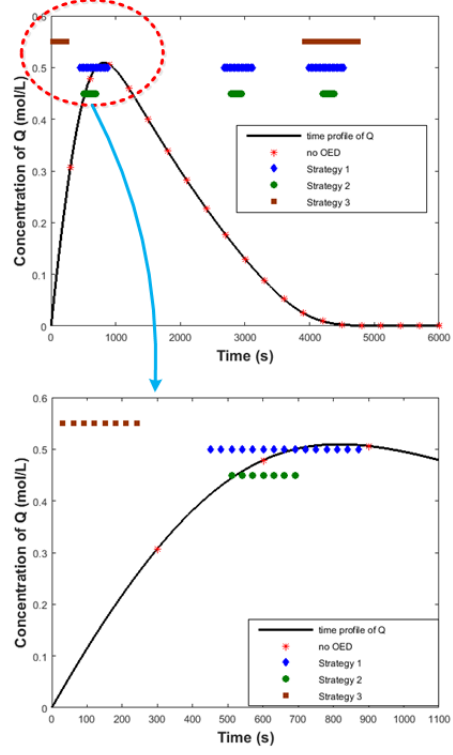
Table 2: Observation design results for enzyme reaction system

Methods	Selected mea- surement states	Sampling time points (unit: sec- ond)
no-OED	{S, P, N, Q, R}	[300:300:6000]
Strategy 1	{S, Q}	[450:30:870], [2670:30:3120], [3390:30:4530]
Strategy 2	{S, P, N, Q, R}	[510:30:690], [2790:30:2940], [4200:30:4380]
Strategy 3	{S}	[420:30:1020], [2130:30:3390]
	{Q}	[30:30:240], [3930:30:4740]

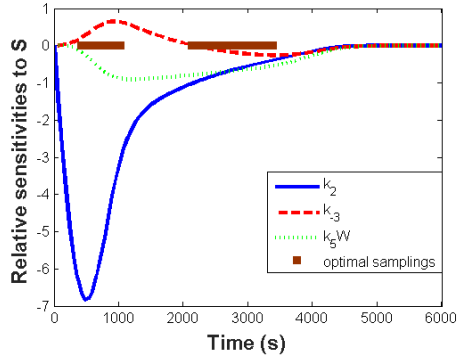
Note: each region of the sampling time profile, in all tables, is shown as
[initial time: sample interval: final time]



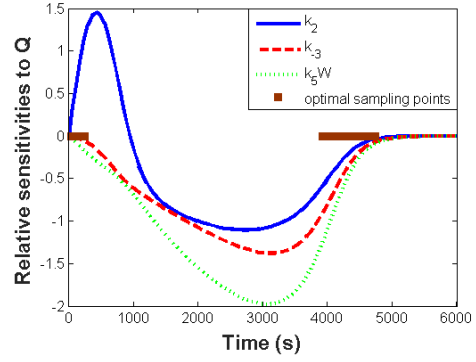
(a) Sampling profile for S



(b) Sampling profile for Q

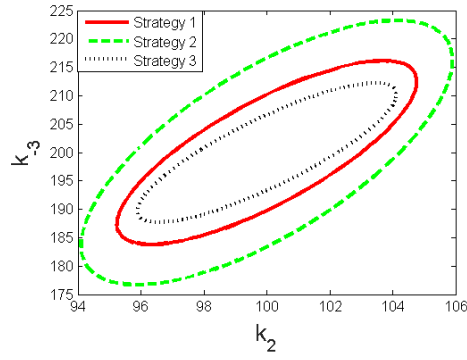


(c) Local sensitivities and sampling of S

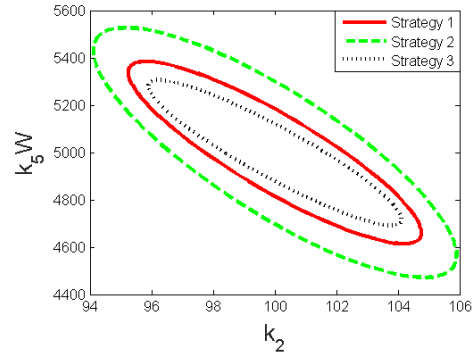


(d) Local sensitivities and sampling of Q

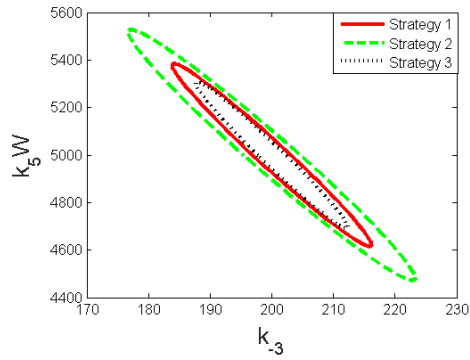
Figure 7: Sampling profiles of S and Q with different OED strategies



(a) Confidence interval of (k_2, k_{-3})



(b) Confidence interval of (k_2, k_5W)



(c) Confidence interval of (k_{-3}, k_5W)

Figure 8: Comparison of confidence intervals with different observation design (enzyme reaction system)

713 *6.1.4. Iterative two-layer design of input and observation*

714 In this section, input and observation variables are designed together
715 through the proposed iterative double-layer OED strategy as shown in Fig.
716 2. The results are compared to another iterative OED, but the observations
717 are designed using Strategy 1, as discussed in Section 6.1.3. In both cases,
718 the iteration number is set to be 100. In both methods, the typical run time
719 of the optimization process is around 1.5 hours on a personal computer with
720 i5-2400 CPU and 4GB memory. The designed results are shown in Table
721 3. By considering both the input and observation factors, S and Q are both
722 selected for the measurement set, which is consistent with the observation
723 design results in Section 6.1.3. The state of N is also found important for
724 measurement set when Strategy 3 is used in the observation design. The CIs
725 of the selected key parameter pairs are shown in Fig. 9. Again, it can be seen
726 that using the same computational time, the results from the proposed OED
727 method provides (potentially) better parameter estimation quality compared
728 with the method with sequential observation design.

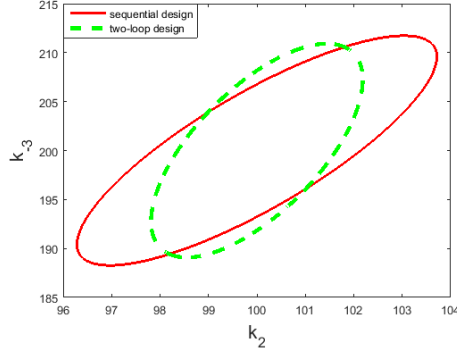
729 One should note that using the proposed observation design or the it-
730 erative double-layer design strategy, non-uniform sampling time regions are
731 suggested to do the measurement rather than the uniform sampling strategy.
732 The latter has been widely used in chemical engineering. Taking uniform
733 sampling at the very early stage of modeling and design would be useful,
734 where model information is limited and parameter values contain large un-
735 certainties.

Table 3: Iterative two-layer OED of input and observation

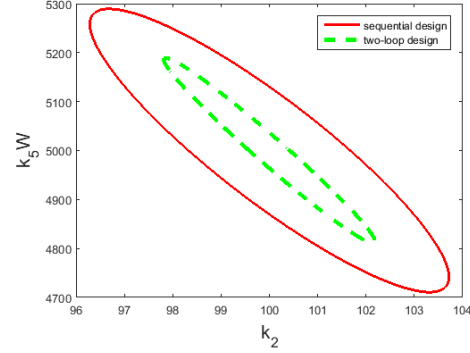
Two-layer OED	$[S_0, E_0, N_0]$: (unit: mol/L)	measur. set	Sampling points (unit: second)
Lower layer: Strategy 1	$[0.74, 1.52e - 5, 1]$	$\{S, Q\}$	$[420:30:1020]$, $[2130:30:3390]$ for S; and $[30:30:240]$, $[3930:30:4740]$ for Q
Lower layer: Strategy 3	$[1, 5.64e - 6, 0.13]$	$\{S, Q, N\}$	$[4590:30:5760]$ for S; $[5280:30:6000]$ for Q; and $[390:30:1410]$ for N

6.2. Enzymatic biodiesel production system

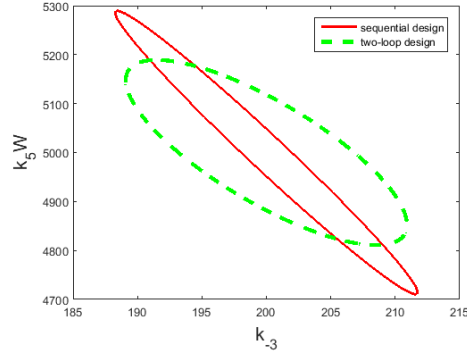
A kinetic model for a lab-scale enzymatic transesterification of rapeseed oil with methanol using Callera Trans L (a liquid formulation of a modified *Thermomyces lanuginosus* lipase) was developed in (Price et al., 2014). In this model, the methanol inhibition and the interfacial and bulk concentrations of the enzyme are considered except for the enzyme deactivation process. The developed model describes the effect of different oil compositions, as well as different water, enzyme and methanol concentrations, which are the relevant conditions required for process evaluation of industrial production of biodiesel. Fig. 10 demonstrates the reaction scheme of this enzymatic biodiesel production system. The free enzyme contained in the polar phase is absorbed at the water oil interface and forms the penetrated enzyme, which further reacts with triglyceride (T), diglyceride (D) and monoglyceride (M) to form enzyme substrate complexes ET, ED and EM. Then these enzyme



(a) Confidence interval of (k_2, k_{-3})



(b) Confidence interval of (k_2, k_5W)



(c) Confidence interval of (k_{-3}, k_5W)

Figure 9: Comparison of confidence intervals under different OEDs of input and observation (enzyme reaction system)

750 substrates can be decomposed into the acyl enzyme complex and D, M and G,
 751 respectively. The acyl enzyme complex can then react with water or methanol
 752 and produce the free fatty acid (FFA) and biodiesel (BD). Additionally, the
 753 competitive methanol inhibition is also considered in this reaction process.
 754 From these kinetic reactions a set of ODEs can be formulated following the
 755 mass-balance principle (Appendix C).

756 A set of experiments have been conducted in advance in order to collect

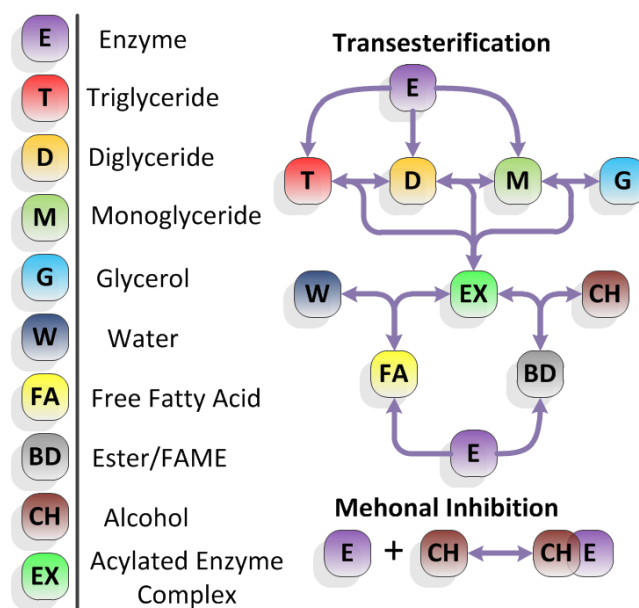


Figure 10: Enzymatic biodiesel production system

757 experimental data for parameter estimation. In all those experiments, the
 758 contents of water and enzyme were varied from 3 to 7 and 0.1 to 0.5 wt.% oil
 759 respectively. An amount of 1.5 equivalents of methanol was reacted with the
 760 Rapeseed oil. One equivalent corresponds to the stoichiometric amount of
 761 alcohol needed to convert all fatty acid residues in the oil to biodiesel. The
 762 reaction was carried out in a 0.25 liter glass reactor with a tank diameter
 763 (T) of 55 mm and 2 baffles, each is $0.18T$ wide. The reactor was immersed
 764 in a water bath with temperature control maintained at 35°C . Initially 0.2
 765 equivalent of methanol was charged with the oil in the reactor. When the
 766 reaction mixture reached the reaction temperature, the amount of water and
 767 enzyme to be used in the experiment was then added to the reactor and
 768 methanol feeding started. The experiment length is set to be 25 hours and

original samplings take place every 15 minutes in the first hour and then once each hour. The unit for all reactant concentrations is in mol/L. The nominal parameter values, initial conditions and feeding rates are provided in Table C.7 and C.8 in Appendix C.

The orthogonalization-based method is applied to rank parameters and examine parameter correlations so as to select the set of most estimable parameters. This method gives consistent results regarding the 10 estimable parameters using the collinearity index. The three most important parameters identified in this analysis are k_6 , k_8 and k_9 (shown in Fig. 11).

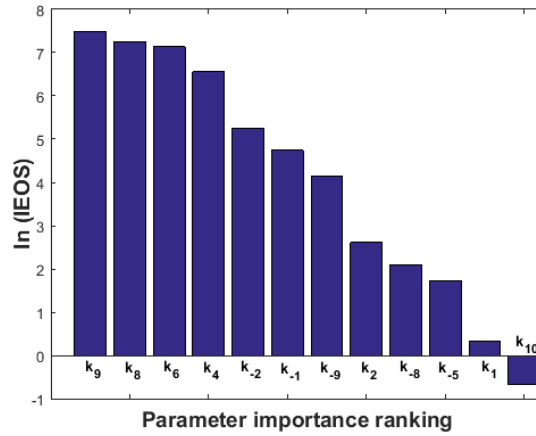


Figure 11: Parameter ranking via orthogonalization (enzymatic biodiesel production system)

Taking the three most important parameters, k_6 , k_8 and k_9 , into the parameter estimation scheme, OED has been applied to determine the best observation strategy which include the most valuable measurement variables and the best sampling time points for each state. Considering the reality

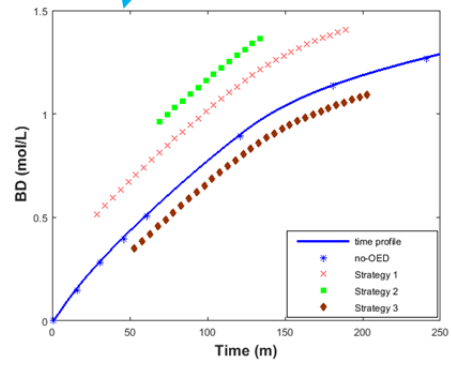
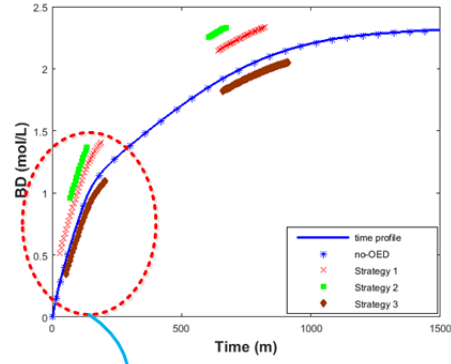
of experimentation, the minimal sampling time interval between two neighboring sampling points is set to be 5 minutes. In non-designed settings, 28 equally spaced sampling points were selected for all five measurable state variables which are T, D, M, BD and FFA. The number of sampling points in this simulation is therefore chosen to be 140 (28×5). Three different experimental strategies in Table 1 are tested, the results of which are shown in Table 4 and in Fig. 12.

Table 4: Design results of different OED strategies

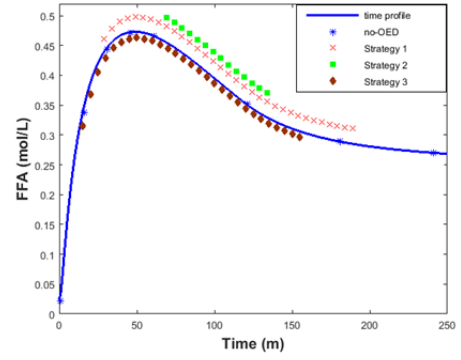
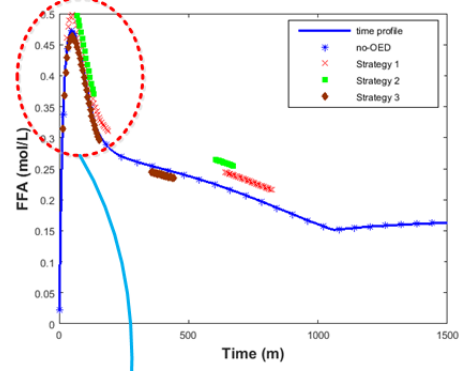
Measurement state variables		Sampling profile (unit: minute)
no-OED	T, D, M, BD, FFA	[0:15:60], [120:60:1440]
Strategy 1	B, FFA	[28:5:188], [640:5:820]
Strategy 2	T, D, M, B, FFA	[68:5:133], [605:5:670]
	M	[121:5:136], [435:5:465]
Strategy 3	B	[52:5:202], [658:5:911]
	FFA	[14:5:154], [356:5:441]

All three OEDs give two sampling regions on the selected or all measurement variables. The measurement should be taken for BD and FFA at the start (first 200 minutes) and middle (between 600 and 1000 minutes) stages of the reaction. This is reasonable because the changes of FFA and M are significant from the start of the reaction. Sampling points selected in this region can grab dynamic information of the system. Also, from Fig. 12(d) it can be seen that the sensitivities of key parameters to BD in the middle reaction stage are quite high. Therefore, additional samplings should be taken in

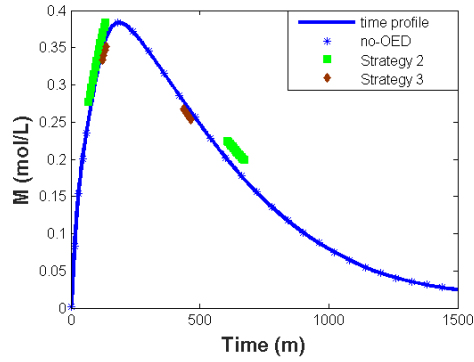
797 this region. Of course the sampling details are not the same when different
798 OEDs are implemented. The sequential designs with Strategy 1 and Strategy
799 2 show that BD and FFA are the most valuable state variables, while in the
800 integrated observation design of Strategy 3, sampling points for M are also
801 shown to be useful. The observation design results are further assessed by
802 comparing the CIs of key parameter pairs in Fig. 13. It can be seen that CIs
803 of all OEDs are smaller than the scenario without OED, and the proposed
804 Strategy 3 achieves the smallest CIs among all OEDs.



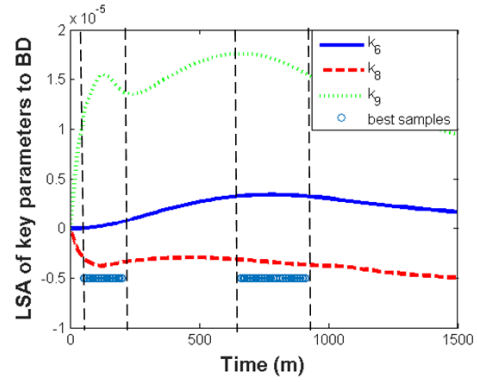
(a) Sampling points for BD



(b) Sampling points for FFA

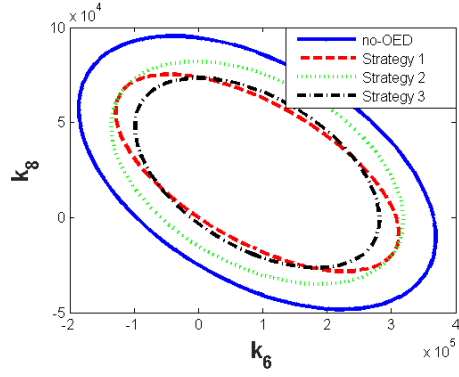


(c) Sampling points for M

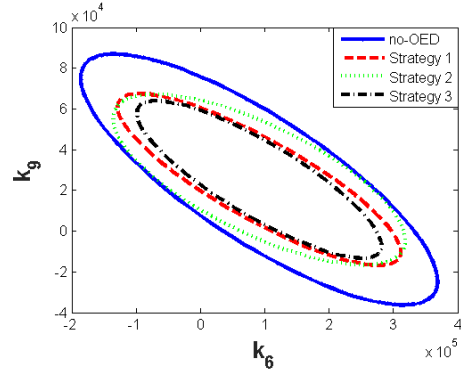


(d) Sampling points and sensitivities of BD

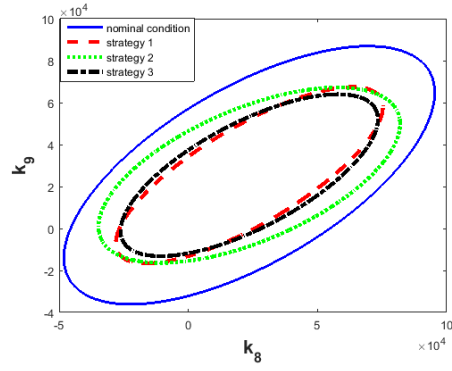
Figure 12: Sampling points on selected state variables (enzymatic biodiesel production system)



(a) Confidence intervals of (k_6, k_8)



(b) Confidence intervals of (k_6, k_9)



(c) Confidence intervals of (k_8, k_9)

Figure 13: Comparison of confidence intervals for different OED strategies (enzymatic biodiesel production system)

805 7. Conclusions and discussions

806 In this work, three experimental design objectives, the input design, the
807 sampling time design and the measurement set selection are investigated.
808 The integrated observation design that determines both measurement set se-
809 lection and sampling time scheduling, simultaneously, has been proposed. By
810 approximating available sampling points *a priori*, the problem formulation
811 for sampling time design can be expressed in a similar form of measurement
812 set selection design. Therefore these two design tasks can be combined to-
813 gether as a single optimization problem, which is further relaxed to a convex
814 optimization problem that can be conveniently solved using local optimiza-
815 tion methods. Furthermore, we have developed an iterative two-layer numer-
816 ical strategy which can deal with OED taking into account input and obser-
817 vation variables together. This new optimization strategy intends to obtain
818 the global optimal results for all experimental conditions in one optimization
819 framework. The input design that is formulated as non-convex optimiza-
820 tion problem is solved by a modern heuristic algorithm, PSO method, in the
821 upper layer. The integrated observation design which can be relaxed into
822 convex optimization problem is solved by a local optimization method, the
823 Powell’s method, in the lower layer. In each iteration, the local optimization
824 and the global optimization are handled separately.

825 Through the case studies based on an enzyme reaction model and a ki-
826 netic model developed for a lab-scale enzyme-catalysed biodiesel production
827 process, the effectiveness of the integrated observation design over two tradi-
828 tional sequential design strategies has been examined. In both case studies,
829 the lower bounds for parameter estimation errors can be reduced through the

830 proposed observation design. Another advantage of this proposed method
831 is that it can automatically choose the number and position of measure-
832 ment points for each measurable state variable rather than measuring all
833 state variables using the same sampling profile. Similar improvement can
834 be observed when the input is included in the iterative two-layer design.
835 The resulted non-uniform sampling time selection is rather non-intuitive but
836 could be of more values compared with conventional uniform sampling sched-
837 ule. Whether the non-uniform sampling schedule is generally appropriate for
838 wider applications need more investigations in future work. It is expected
839 that a well-designed sampling schedule contains more useful information and
840 a non-uniform sampling should be more cost effective compared with con-
841 ventional uniform sampling.

842 OED is a model-based technology, the results of which depend on the
843 prior knowledge of the system model, also on the design criteria and the op-
844 timization methods used. It is therefore not always possible to get consistent
845 OED results under various circumstances. Nevertheless, this is a systematic
846 method that can provide useful guidance to experimental settings, with the
847 benefits of collecting measurement data that are most valuable to process
848 modeling. The OED results can sometimes be different from the experi-
849 ences or intuitive understanding of the experimental conditions, for example,
850 key regions in sampling rather than uniform sampling for measurement data
851 can be revealed by OED, which may not be obvious from the experimen-
852 tal practice. The measurement set selection may suggest useful variables
853 that are ignored in the existing measurement system. Further development
854 on model-based OED methodology can be investigated by considering more

855 complicated factors required by applications, for example, model uncertain-
856 ties during the design stage, non-Gaussian noise in measurement, design of
857 time-dependent experimental factors. All these tasks will be very challeng-
858 ing.

859 **Acknowledgment**

860 The authors would like to thank Dr Jason Price and his colleagues from
861 the Department of Chemical and Biochemical Engineering, Technical Uni-
862 versity of Denmark, for providing the mathematical model of the enzymatic
863 biodiesel production system and for many useful discussions.

864 **Bibliography**

- 865 Asyali, M. H., 2010. Design of optimal sampling times for pharmacokinetic
866 trials via spline approximation. *Turkish Journal of Electrical Engineering*
867 & *Computer Sciences* 18 (6), 1019–1030.
- 868 Atherton, R., Schainker, R., Ducot, E., 1975. On the statistical sensitivity
869 analysis of models for chemical kinetics. *AIChE Journal* 21 (3), 441–448.
- 870 Balsa-Canto, E., Alonso, A. A., Banga, J. R., 2008. Computational pro-
871 cedures for optimal experimental design in biological systems. *Systems*
872 *Biology, IET* 2 (4), 163–172.
- 873 Balsa-Canto, E., Alonso, A. A., Banga, J. R., 2010. An iterative identification
874 procedure for dynamic modeling of biochemical networks. *BMC Systems*
875 *Biology* 4 (1), 11–28.

- 876 Baltes, M., Schneider, R., Sturm, C., Reuss, M., 1994. Optimal experimental
877 design for parameter estimation in unstructured growth models. *Biotech-*
878 *nology Progress* 10 (5), 480–488.
- 879 Banga, J. R., Balsa-Canto, E., Moles, C. G., Alonso, A. A., 2005. Dynamic
880 optimization of bioprocesses: Efficient and robust numerical strategies.
881 *Journal of Biotechnology* 117 (4), 407–419.
- 882 Banga, J. R., Versyck, K. J., Van Impe, J. F., 2002. Computation of op-
883 timal identification experiments for nonlinear dynamic process models: a
884 stochastic global optimization approach. *Industrial & Engineering Chem-*
885 *istry Research* 41 (10), 2425–2430.
- 886 Ben-Zvi, A., McLellan, P. J., McAuley, K., 2006. Identifiability of non-linear
887 differential algebraic systems via a linearization approach. *The Canadian*
888 *Journal of Chemical Engineering* 84 (5), 590–596.
- 889 Biegler, L. T., Cervantes, A. M., Wächter, A., 2002. Advances in simulta-
890 neous strategies for dynamic process optimization. *Chemical Engineering*
891 *Science* 57 (4), 575–593.
- 892 Bogacka, B., Patan, M., Johnson, P. J., Youdim, K., Atkinson, A. C., 2011.
893 Optimum design of experiments for enzyme inhibition kinetic models. *Jour-*
894 *nal of Biopharmaceutical Statistics* 21 (3), 555–572.
- 895 Brown, M., He, F., Yeung, L. F., 2008. Robust measurement selection for
896 biochemical pathway experimental design. *International Journal of Bioin-*
897 *formatics Research and Applications* 4 (4), 400–416.

- 898 Brun, R., Reichert, P., Künsch, H. R., 2001. Practical identifiability analysis
899 of large environmental simulation models. *Water Resources Research*
900 37 (4), 1015–1030.
- 901 Catania, F., Paladino, O., 2009. Optimal sampling for the estimation of
902 dispersion parameters in soil columns using an iterative genetic algorithm.
903 *Environmental Modelling & Software* 24 (1), 115–123.
- 904 Chianeh, H. A., Stigter, J., Keesman, K. J., 2011. Optimal input design
905 for parameter estimation in a single and double tank system through direct
906 control of parametric output sensitivities. *Journal of Process Control*
907 21 (1), 111–118.
- 908 Chis, O.-T., Banga, J. R., Balsa-Canto, E., 2011. Structural identifiability of
909 systems biology models: a critical comparison of methods. *PloS one* 6 (11),
910 1–16.
- 911 de Brauwere, A., De Ridder, F., Gourgue, O., Lambrechts, J., Comblen,
912 R., Pintelon, R., Passerat, J., Servais, P., Elskens, M., Baeyens, W., et al.,
913 2009. Design of a sampling strategy to optimally calibrate a reactive transport
914 model: Exploring the potential for escherichia coli in the scheldt estuary.
915 *Environmental Modelling & Software* 24 (8), 969–981.
- 916 Fages, F., Soliman, S., Chabrier-Rivier, N., 2004. Modelling and querying
917 interaction networks in the biochemical abstract machine biocham. *Journal*
918 *of Biological Physics and Chemistry* 4 (1), 64–73.
- 919 Faller, D., Klingmüller, U., Timmer, J., 2003. Simulation methods for optimal
920 experimental design in systems biology. *Simulation* 79 (12), 717–725.

921 Flaherty, P., Jordan, M., Arkin, A., et al., 2006. Robust design of biological
922 experiments. *Advances in Neural Information Processing Systems* 18, 363–
923 370.

924 Fletcher, R., Powell, M. J., 1963. A rapidly convergent descent method for
925 minimization. *The computer journal* 6 (2), 163–168.

926 Franceschini, G., Macchietto, S., 2008. Model-based design of experiments
927 for parameter precision: State of the art. *Chemical Engineering Science*
928 63 (19), 4846–4872.

929 He, F., Brown, M., Yue, H., 2010. Maximin and bayesian robust experimental
930 design for measurement set selection in modelling biochemical regulatory
931 systems. *International Journal of Robust and Nonlinear Control* 20 (9),
932 1059–1078.

933 Kennedy, J., 2011. Particle swarm optimization. In: *Encyclopedia of machine*
934 *learning*. Springer, pp. 760–766.

935 Kutalik, Z., Cho, K.-H., Wolkenhauer, O., 2004. Optimal sampling time se-
936 lection for parameter estimation in dynamic pathway modeling. *Biosystems*
937 75 (1), 43–55.

938 Liepe, J., Filippi, S., Komorowski, M., Stumpf, M. P. H., 01 2013. Maxi-
939 mizing the information content of experiments in systems biology. *PLOS*
940 *Computational Biology* 9 (1), 1–13.

941 Ljung, L., 1987. *System identification: theory for the user*. NJ: Prentice
942 Hall.

- 943 Ljung, L., Glad, T., 1994. On global identifiability for arbitrary model
944 parametrizations. *Automatica* 30 (2), 265–276.
- 945 McLean, K. A., McAuley, K. B., 2012. Mathematical modelling of chemical
946 processes obtaining the best model predictions and parameter estimates
947 using identifiability and estimability procedures. *The Canadian Journal of*
948 *Chemical Engineering* 90 (2), 351–366.
- 949 Peleg, M., Yeh, I., Altman, R. B., 2002. Modelling biological processes using
950 workflow and petri net models. *Bioinformatics* 18 (6), 825–837.
- 951 Phair, R. D., 1997. Development of kinetic models in the nonlinear world of
952 molecular cell biology. *Metabolism* 46 (12), 1489–1495.
- 953 Pohjanpalo, H., 1978. System identifiability based on the power series expan-
954 sion of the solution. *Mathematical Biosciences* 41 (1), 21–33.
- 955 Price, J., Hofmann, B., Silva, V. T., Nordblad, M., Woodley, J. M., Huusom,
956 J. K., 2014. Mechanistic modeling of biodiesel production using a liquid
957 lipase formulation. *Biotechnology Progress* 30 (6), 1277–1290.
- 958 Ruffio, E., Saury, D., Petit, D., 2012. Robust experiment design for the
959 estimation of thermophysical parameters using stochastic algorithms. *In-*
960 *ternational Journal of Heat and Mass Transfer* 55 (11), 2901–2915.
- 961 Sandink, C., McAuley, K., McLellan, P., 2001. Selection of parameters for
962 updating in on-line models. *Industrial & Engineering Chemistry Research*
963 40 (18), 3936–3950.

- 964 Sun, C., Hahn, J., 2006. Parameter reduction for stable dynamical systems
965 based on hankel singular values and sensitivity analysis. *Chemical Engi-
966 neering Science* 61 (16), 5393–5403.
- 967 Vajda, S., Godfrey, K. R., Rabitz, H., 1989. Similarity transformation
968 approach to identifiability analysis of nonlinear compartmental models.
969 *Mathematical Biosciences* 93 (2), 217–248.
- 970 van Riel, N. A., 2006. Dynamic modelling and analysis of biochemical net-
971 works: mechanism-based models and model-based experiments. *Briefings
972 in Bioinformatics* 7 (4), 364–374.
- 973 Villaverde, A. F., Henriques, D., Smallbone, K., Bongard, S., Schmid, J.,
974 Cicin-Sain, D., Crombach, A., Saez-Rodriguez, J., Mauch, K., Balsa-
975 Canto, E., et al., 2014. Biopredyn-bench: benchmark problems for kinetic
976 modelling in systems biology. *arXiv preprint arXiv:1407.5856*.
- 977 Walter, E., Lecourtier, Y., 1982. Global approaches to identifiability testing
978 for linear and nonlinear state space models. *Mathematics and Computers
979 in Simulation* 24 (6), 472–482.
- 980 Yao, K. Z., Shaw, B. M., Kou, B., McAuley, K. B., Bacon, D., 2003. Modeling
981 ethylene/butene copolymerization with multi-site catalysts: parameter es-
982 timability and experimental design. *Polymer Reaction Engineering* 11 (3),
983 563–588.
- 984 Yu, H., Yue, H., Halling, P., 2015. Optimal experimental design for an enzy-
985 matic biodiesel production system. *IFAC-PapersOnLine* 48 (8), 1258–1263.

986 Yue, H., Brown, M., He, F., Jia, J., Kell, D. B., 2008. Sensitivity analysis
 987 and robust experimental design of a signal transduction pathway system.
 988 International Journal of Chemical Kinetics 40 (11), 730–741.

989 Yue, H., Halling, P., Yu, H., 2013. Model development and optimal experi-
 990 mental design of a kinetically controlled synthesis system. IFAC Proceed-
 991 ings Volumes 46 (31), 327–332.

992 **Appendix A. Orthogonalized sensitivity analysis**

993 The basic step of orthogonalization based forward selection method is
 994 described as follows.

- 995 1. The normalized parameter sensitivity $\bar{\mathbf{S}}$ and the magnitude of each
 996 column in $\bar{\mathbf{S}}$ is calculated based on (13) and (3), respectively. The
 997 parameter corresponding to the column with maximum magnitude is
 998 the first identifiable parameter. Set $k=1$.
- 999 2. Put the k columns from $\bar{\mathbf{S}}$ that correspond to parameters that have
 1000 been identified into matrix \mathbf{X}_k .
- 1001 3. Use \mathbf{X}_k to calculate the ordinary least-square prediction of matrix $\bar{\mathbf{S}}$:

$$\hat{\mathbf{S}}_k = \mathbf{X}_k (\mathbf{X}_k^T \cdot \mathbf{X}_k)^{-1} \mathbf{X}_k \bar{\mathbf{S}}$$

 1002 and calculate the residual matrix by $\mathbf{R}_k = \bar{\mathbf{S}} - \hat{\mathbf{S}}_k$.
- 1003 4. Calculate the magnitude of each column in \mathbf{R}_k . The $(k + 1)$ -th most
 1004 identifiable parameter corresponds to the column in \mathbf{R}_k with the largest
 1005 magnitude.
- 1006 5. Increase k by 1 and add the column of $\bar{\mathbf{S}}$ that corresponds to the $(k+1)$ -
 1007 th parameter to matrix \mathbf{X}_k .

1008 6. Repeat steps 3-5 for all parameters until the maximum magnitude in
 1009 \mathbf{R}_k is less than a predefined threshold.

1010 **Appendix B. Supplementary materials of the enzyme reaction sys-** 1011 **tem**

1012 The 10 ordinary differential equations for the enzyme reaction system are
 1013 as follows.

$$\begin{aligned} \frac{dE}{dt} = & -k_1 \cdot E \cdot S + k_{-1} \cdot ES + k_4 \cdot EQ - k_{-4} \cdot E \cdot Q \\ & + k_6 \cdot ER \end{aligned} \quad (\text{B.1})$$

$$\frac{dES}{dt} = k_1 \cdot E \cdot S - k_{-1} \cdot ES - k_2 \cdot ES + k_{-2} \cdot E^* \cdot P \quad (\text{B.2})$$

$$\begin{aligned} \frac{dE^*}{dt} = & k_2 \cdot ES - k_{-2} \cdot E^* \cdot P - k_3 \cdot E^* \cdot N + k_{-3} \cdot EQ \\ & - k_5 \cdot W \cdot E^* + k_{-5} \cdot ER \end{aligned} \quad (\text{B.3})$$

$$\frac{dEQ}{dt} = k_3 \cdot E^* \cdot N - k_{-3} \cdot EQ - k_4 \cdot EQ + k_{-4} \cdot E \cdot Q \quad (\text{B.4})$$

$$\frac{dER}{dt} = k_5 \cdot W \cdot E^* - k_{-5} \cdot ER - k_6 \cdot ER \quad (\text{B.5})$$

$$\frac{dS}{dt} = -K_1 \cdot E \cdot S + k_{-1} \cdot ES \quad (\text{B.6})$$

$$\frac{dP}{dt} = k_2 \cdot ES - k_{-2} \cdot E^* \cdot P \quad (\text{B.7})$$

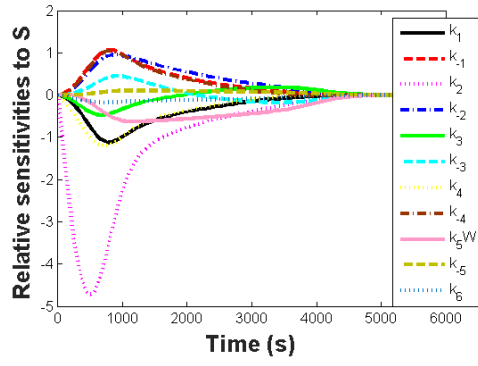
$$\frac{dN}{dt} = -k_3 \cdot E^* \cdot N + k_{-3} \cdot EQ \quad (\text{B.8})$$

$$\frac{dQ}{dt} = k_4 \cdot EQ - k_{-4} \cdot E \cdot Q \quad (\text{B.9})$$

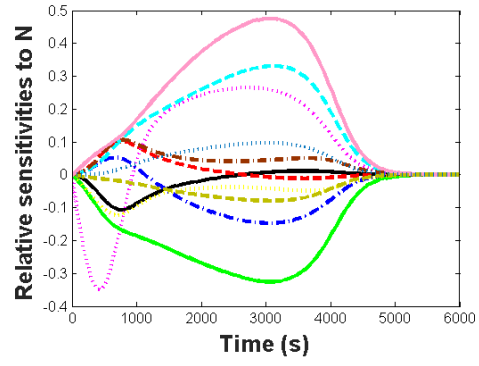
$$\frac{dR}{dt} = k_6 \cdot ER \quad (\text{B.10})$$

Table B.5: List of state variables and kinetic parameters

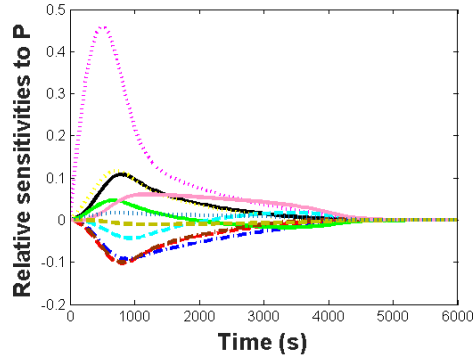
State variables	Initial condition ($mol \cdot L^{-1}$)	Kinetic parameters	Nominal values
$S(x_1)$	0.8	k_1	1e5
$P(x_2)$	0	k_{-1}	1e3
$N(x_3)$	0.9	k_2	100
$Q(x_4)$	0	k_{-2}	1e4
$R(x_5)$	0	k_3	5e4
$E(x_6)$	1.5e-5	k_{-3}	200
$E^*(x_7)$	0	k_4	1e3
$ES(x_8)$	0	k_{-4}	2e4
$EQ(x_9)$	0	$k_5 W$	5e3
$ER(x_{10})$	0	k_{-5}	100
		k_6	500



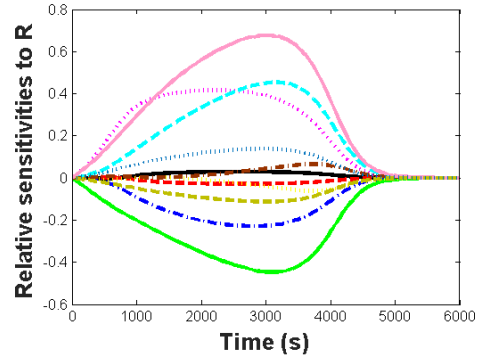
(a) Relative sensitivities to S



(b) Relative sensitivities to N

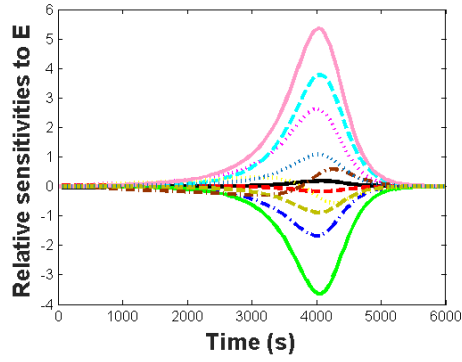


(c) Relative sensitivities to P

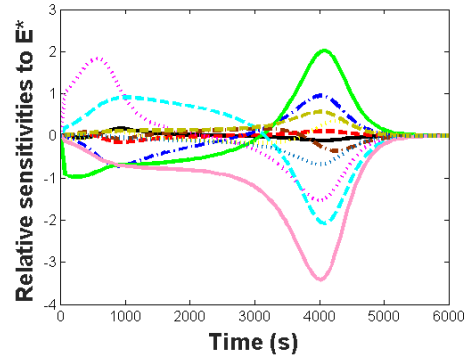


(d) Relative sensitivities to R

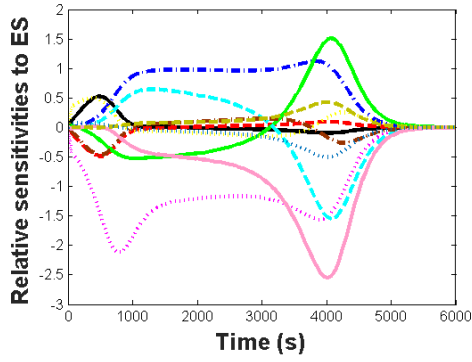
Figure B.14: Parameter relative sensitivities to S, N, P, R (enzyme reaction system)



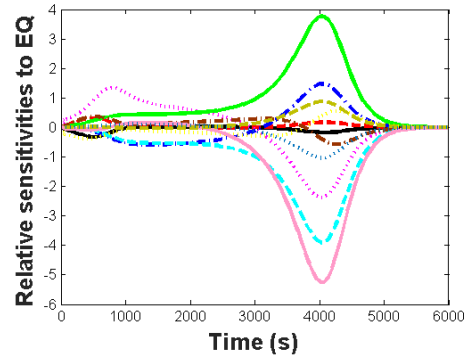
(a) Relative sensitivities to E



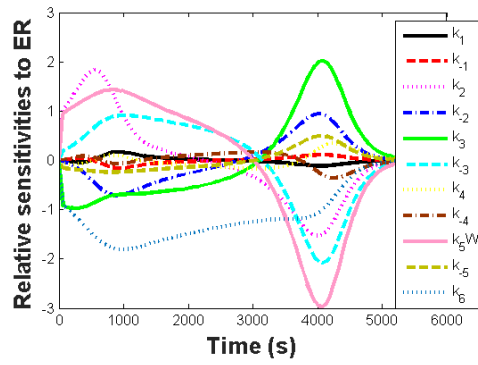
(b) Relative sensitivities to E^*



(c) Relative sensitivities to ES

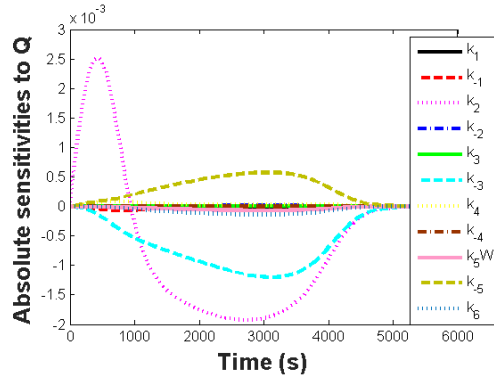


(d) Relative sensitivities to EQ

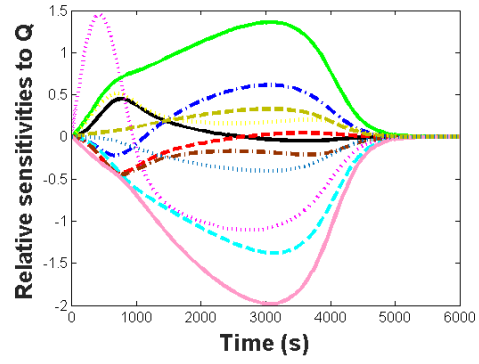


(e) Relative sensitivities to ER

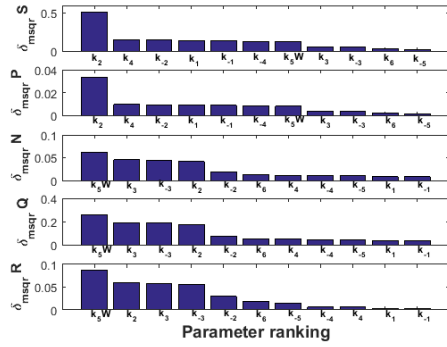
Figure B.15: Parameter relative sensitivities to non-measurable enzyme complexes (enzyme reaction system)



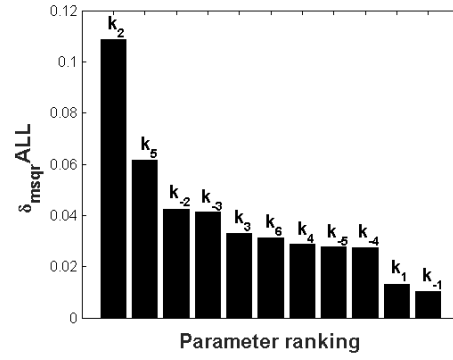
(a) Absolute sensitivities to Q



(b) Relative sensitivities to Q



(c) LSA parameter ranking for each state



(d) LSA parameter ranking for all states

Figure B.16: Parameter ranking for enzyme reaction system

1014 **Appendix C. Supplementary information of enzymatic biodiesel**
1015 **production system**

1016 Ordinary differential equations of enzymatic biodiesel production system:

$$\frac{d([T] \cdot V)}{dt} = -V(r_2) \quad (C.1)$$

$$\frac{d([D] \cdot V)}{dt} = V(r_3 - r_4) \quad (C.2)$$

$$\frac{d([M] \cdot V)}{dt} = V(r_5 - r_6) \quad (C.3)$$

$$\frac{d([BD] \cdot V)}{dt} = V(r_9) \quad (C.4)$$

$$\frac{d([FA] \cdot V)}{dt} = V(r_8) \quad (C.5)$$

$$\frac{d([G] \cdot V)}{dt} = V(r_7) \quad (C.6)$$

$$\frac{d([W] \cdot V)}{dt} = -V(r_8) \quad (C.7)$$

$$\frac{d([CH] \cdot V)}{dt} = -V(r_9 + r_{10}) \quad (C.8)$$

$$\frac{d([E] \cdot V)}{dt} = V(r_1 + r_8 + r_9 - r_2 - r_4 - r_6 - r_{10}) \quad (C.9)$$

$$\frac{d([EX] \cdot V)}{dt} = V(r_3 + r_5 + r_7 - r_8 - r_9) \quad (C.10)$$

$$\frac{d([ET] \cdot V)}{dt} = V(r_2 - r_3) \quad (C.11)$$

$$\frac{d([ED] \cdot V)}{dt} = V(r_5 - r_6) \quad (C.12)$$

$$\frac{d([EM] \cdot V)}{dt} = V(r_6 - r_7) \quad (C.13)$$

$$\frac{d([ECH] \cdot V)}{dt} = V(r_{10}) \quad (C.14)$$

$$\frac{d([E_{bulk}] \cdot V)}{dt} = -V(r_1) \quad (C.15)$$

$$\frac{d(V_p)}{dt} = R_G + R_W \quad (C.16)$$

$$\frac{d(V)}{dt} = (F_a) \quad (C.17)$$

Table C.6: Kinetic mechanism for the enzymatic transesterification

$E_{bulk} + A_f \leftrightarrow E$	$r_1 = k_1 \cdot E_{bulk} \cdot A_f - k_{-1} \cdot E$
$T + E \leftrightarrow ET$	$r_2 = k_2 \cdot T \cdot E - k_{-2} \cdot ET$
$ET \leftrightarrow EX + D$	$r_3 = k_3 \cdot ET - k_{-3} \cdot EX \cdot D$
$D + E \leftrightarrow ED$	$r_4 = k_4 \cdot D \cdot E - k_{-4} \cdot ED$
$ED \leftrightarrow EX + M$	$r_5 = k_5 \cdot ED - k_{-5} \cdot EX \cdot M$
$M + E \leftrightarrow EM$	$r_6 = k_6 \cdot M \cdot E - k_{-6} \cdot EM$
$EM \leftrightarrow EX + G$	$r_7 = k_7 \cdot EM - k_{-7} \cdot EX \cdot G$
$EX + W \leftrightarrow FA + E$	$r_8 = k_8 \cdot EX \cdot W - k_{-8} \cdot FA \cdot E$
$EX + CH \leftrightarrow BD + E$	$r_9 = k_9 \cdot EX \cdot CH - k_{-9} \cdot BD \cdot E$
$CH + E \leftrightarrow ECH$	$r_{10} = k_{10} \cdot CH \cdot E - k_{-10} \cdot ECH$

Table C.7: Nominal parameter values for enzyme biodiesel production system

k_1	4.95e4	k_6	9.13e4
k_{-1}	6.60	k_{-6}	5.43e5
k_2	1.69e6	k_7	7.06e6
k_{-2}	1.11e4	k_{-7}	4.93
k_3	2.07e4	k_8	2.36e4
k_{-3}	2.20e7	k_{-8}	3.51e6
k_4	3.41e6	k_9	2.54e4
k_{-4}	1.33e7	k_{-9}	2.05e5
k_5	1.55e7	k_{10}	3.23e-2
k_{-5}	1.81e5	k_{-10}	4.39e-4

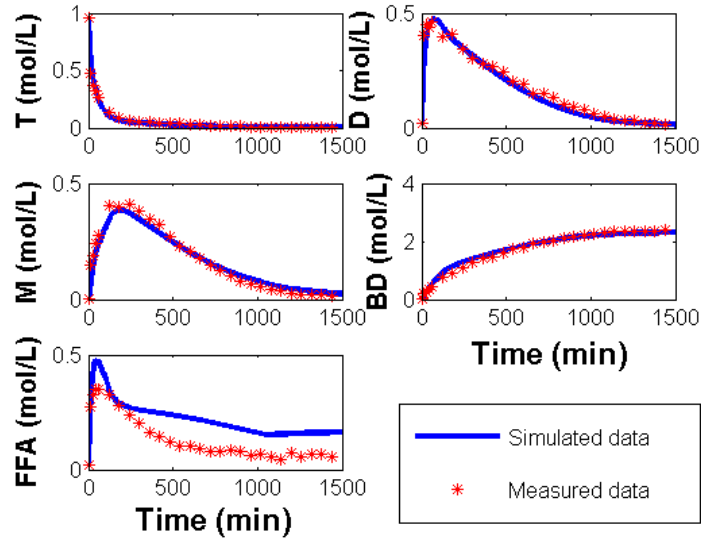


Figure C.17: Time profile of state variables for enzymatic biodiesel production system

Table C.8: Initial input values and feeding rate of methanol

Species	Ini. cond. ($mol \cdot L^{-1}$)	Species	Ini. cond. ($mol \cdot L^{-1}$)
$T(x_1)$	0.9536	$EX(x_{10})$	0
$D(x_2)$	0.0195	$ET(x_{11})$	0
$M(x_3)$	0.0014	$ED(x_{12})$	0
$B(x_4)$	1e-4	$EM(x_{13})$	0
$FFA(x_5)$	0.0224	$ECH(x_{14})$	0
$G(x_6)$	1e-6	$Ef(x_{15})$	9.7165e-6
$W(x_7)$	2.3854	$Vp(x_{16})$	0.0661
$CH(x_8)$	0.5850	$V(x_{17})$	1.5383
$E(x_9)$	0		
Methanol feed rate [$eq \cdot h^{-1}$]	Initial dose methanol [eq]	water [wt.% oil]	Enzyme [wt.% oil]
0.185 first 2hrs; 0.06 thereafter	0.2	5	0.5